DOCUMENT RESUME

ED 083 159 SP 006 945

AUTHOR	Menges, Robert J.
TITLE	Assessing Readiness for Professional Practice. Occasional Paper Number One.
INSTITUTION	North Western Univ., Evanston, Ill. Center for the Teaching Professions.
SPONS AGENCY	American Association of Theological Schools, Dayton, Ohio.
PUB DATE	Jul 73
NOTE	67p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Certification; *Credentials; *Professional Education; *Professional Occupations; Professional Personnel; *Standards; Teachers

ABSTRACT

This report asks what are the characteristics a person should possess before being admitted to practice as a professional, how and by whom are those characteristics identified, and how are they measured. It identifies and selects certain criteria common to all professionals; then it details various aspects involved in assessment and certification. Some of these aspects are personality, knowledge and ability to apply subject matter, and job performance. A list of resource journals and major associations and agencies contacted and a description of procedures are appended. (JB)

ASSESSING READINESS FOR PROFESSIONAL PRACTICE

Robert J. Menges

Occasional Paper Number One

July, 1973

Acknowledgements

This paper was commissioned by the American Association of Theological Schools as part of The Readiness for Ministry Project. I am indebted to the Association staff for the concept "readiness" as used here. Particularly helpful contributions were made by James Deneen, Alan Knox, and LeRoy Nattress, Jr. Whatever efficiency the project had is due in large part to the research, writing and typing skills of Emily Demme.

Robert J. Menges
Program Director
The Center for the Teaching Professions

Contents

Prologue:

The Lady in the Blue Room

Once upon a time the Royal Court's Undersecretary for Carnal Knowledge called at a House of Prostitution. Explaining that he was acting as agent for the King, he inquired after the merits of the four ladies who lived therein. "They are of equal beauty and of equal price," he was told. "Come, spend some time with each so you will choose only the best for the King."

The Undersecretary was well acquainted with the King's preferences in such matters, for the King never tired of discussing his experiences in such detail that even Undersecretaries could understand.

First, he called on the lady in the yellow room. She gave a charming and sincere description of her background--she came from a long line of prostitutes--saying that it is a profession in which she has always been interested. Indeed, to this day some of her best friends are prostitutes. A psychologist client who gave her several psychological tests told her she has the same personality traits as the others in her profession whom he had tested. She received consistently favorable ratings from the Head of the House and from her coworkers. The visitor thought that she surely would be a success.

Next, he visited the lady in the red room, who showed the same qualities as the first. She delivered a rich discourse in prose and verse, in song and story, on the history of her profession, statistics on its present thriving state, and particularly its success at excluding those who are not worthy to be members. She demonstrated her vast and thorough knowledge of the principles of anatomy and gynecology, and showed she was well-versed on the latest methods of birth control. (Her certificates of distinction for several correspondence courses in those areas were prominently displayed.) She pointed out several shelves of volumes on sex techniques and on the psychology of sexuality and offered incisive answers to questions on the few volumes with which the visitor was familiar.

He then spent some time with the lady in the green room. She said much the same as the first two. She talked further about her experiences and those of her fellow professionals. She told of the most frequent services she is called upon to provide, and discribed a number of typical clients. Her discussion was filled with details of means and modes and standard (as well as not-so-standard) deviations. They then took a tour of the premises during which she pointed out the function of some rather remarkable equipment. She invited questions from him about what she would do in various situations and, never at a loss, responded with candor and a degree of detail that nearly cost the visitor his detachment.

Finally, he was received by the lady in the blue room, where he heard again what had gone before and still more. He was presented there with

several possibilities and invited to accept as many as he wished: to view films of her last half-dozen clients (their identities of course anonymous); to examine her file of letters from previous clients; to remain and observe her next appointment (a young man in therapy who was used to being observed). She even began a disconcerting demonstration followed by an offer that he could hardly refuse.

But because he was a loyal Undersecretary, he merely thanked each of the women and returned to the King deep in thought. The lady in the blue room was his recommendation. It was the start of a beautiful professional relationship.

* * *

This report deals with readiness for professional practice. What are the characteristics (personality traits, knowledge, skills, attitudes, and beliefs) which a person should possess before being permitted to practice as a professional? How and by whom are those characteristics identified? How are they measured?

Any licensing or certification system is less than perfect. That is, some persons will be approved who are actually incompetent and some who are competent are likely to be excluded. It is my position that multiple levels of assessment are likely to maximize selection effectiveness. The lady in the blue room presented more varied evidence than did the others. She showed a) personality traits like those of others in her profession, b) mastery of information presumably related to the practice of her profession, c) the ability to apply subject matter to hypothetical situations, and d) evidence of successful performance on the job. Since more information is available about her than about the others, the prediction of her performance effectiveness is likely to be greater than for any of the others.

We shall see, however, that professional schools and certification and licensing boards seldom gather such extensive information prior to deciding whether a candidate is to be approved. Further, the methods used to gather information are unlikely to have been validated for this purpose, i.e. shown to differentiate effective from ineffective practitioners.

I. Professionals Old and New

The purpose of this chapter is three-fold: to set forth characteristics which justify calling an occupation a profession; to differentiate three categories of credentials pertinent to professions, i. e. accreditation, licensure, and certification or registration; and to summarize several critiques of established professions, with special emphasis on the helping professions.

Criteria for the Professions[1]

Many have wished to benefit from the positive connotations of the term "professional." The literature contains discussions for almost any profession on whether it actually deserves that appellation. For example, in a paper published in 1915, Abraham Flexner asks, "Is Social Work a Profession?". In his discussion he offers six criteria by which a profession may be differentiated from other occupations: it is basically intellectual; it is learned, that is, derived from a set of principles; it is more practical than theoretical; it has a technique, thereby justifying "professional" education; it is strongly organized internally; and its motives include altruism.

Most subsequent writers have endorsed these assertions. Some have suggested additional criteria, two of which seem appropriate to discuss here. First, the role of internal self-discipline, usually administered through professional associations, has been emphasized. Members of the profession are the only persons sufficiently well informed to make disciplinary decisions, according to this argument. Furthermore, the autonomy of the profession is threatened if disciplinary authority is given to external sources.

Second, professional success is to be defined by the quality of service rendered. Other measures of success such as the monetary are therefore said to be of less importance. More recent writings have specified the client's interests as taking precedence over the professional's interests whenever a conflict arises. Quoted apparently with approval in a report on legal education are the words, "Because the clientele is vulnerable the professional owes his first duty to his client's best interests," (Packer and Ehrlich, 1972, p. 22). The current movement in education for teacher accountability proclaims that changes in the student are the data by which teacher success is to be evaluated. We shall return to this contention in our discussions of criteria selection and performance evaluation. Before leaving it, however, one of its implications discussed in some detail by Schein (1972) should be noted. Schein draws a distinction among three client types:

---

[1] This section is based largely on Becker, 1962.

immediate, intermediate, and ultimate. To use an illustration from medical research, suppose a researcher conducts an evaluation of a new drug for the company which desires to market it. The resulting product is to be used by physicians with their patients. From the researcher's point of view, the company is the immediate client, the physician is the intermediate client, and the patient is the ultimate client. If the researcher's professional effectiveness is defined by client impact and if the needs of those three client categories conflict (as well they might), which client is the client? From the drug company's point of view there are at least two client categories, the physician (immediate) and the patient (ultimate), and their interests may also be in conflict. Thus, assessment of effectiveness according to client impact adds as much complexity as clarity to the list of criteria for professions.

Credentials Necessary for Professional Practice

Power to bestow credentials for professional practice is exercised by educational institutions, professional associations, and government agencies. Accreditation is the outcome of evaluating the standards of an educational institution or program. Since such evaluation is typically done by professional associations and government agencies, educational institutions are the least autonomous of the credential sources. Licensure, specifically in the health professions, is "the process by which an agency of government grants permission to persons to engage in a given profession or occupation by certifying that those licensed have attained the minimal degree of competency necessary to ensure that the public health, safety, and welfare will be reasonably well protected," (United States Department of Health, Education, and Welfare, 1971, p. 7). Finally, certification or registration is given by an association or agency to those meeting several qualifications, usually including graduation from an accredited program, passing standard profession-wide examinations, and completing specified supervised work experience.

This report deals primarily with credentials granted to individuals. Therefore, it is concerned least with accreditation since accreditation standards adopted by professional associations deal with programs rather than with individuals. Even though such programs are intended to have quite specific impacts, the standards usually are framed in terms of situations to which students are exposed, rather than in terms of the changes which the experiences produce in those students (c.f. National Council for Accreditation of Teacher Education, 1970).

Licensure is also a secondary concern here primarily because this project has not had resources to study such procedures extensively. My impression is that legislatively determined licensing criteria are seldom derived from a scholarly research base, except as such data are supplied by professional associations. Instead, intuition and biases of legislators interact with power politics to determine licensing standards.

Such power struggles sometimes create odd teammates. For example, in New York state the demand for counseling and psychotherapy services has been greater than the resources of doctoral level practitioners. Because existing law does not restrict the use of titles such as "psychotherapist," "psychoanalyst," and "marriage counselor," these titles have been appropriated by those lacking doctoral qualifications. Legislation (the Biondo bill) was introduced in 1972 to restrict control over such services to Ph.D. psychologists, M.D.'s, M.S.W.'s, and R.N.'s. United in support of the bill with the licensing agency (the State Department of Education) was the unprecedented coalition of the State Attorney General's Office, the State Department of Mental Hygiene, and the state professional associations of the four major mental health professions (psychology, psychiatry, social work, and nursing). In opposition, two coalitions representing psychoanalytic institutes, growth centers, marriage counselors, and paraprofessionals introduced their own bills. It appears unlikely that their bills will pass, but their opposition was sufficient to cause the Biondo bill to be withdrawn in 1973, at the brink of defeat (Asher, 1973). Whatever the result, it will not be based on solid theory or data.

The New Face of Professionalism

Criticising the power of professionals is not a recent pastime and probably began when professionals first associated. George Bernard Shaw is supposed to have observed that all professions are conspiracies against the laity. The late 1960's and early 1970's have seen younger professionals direct especially strong criticism toward their elders in the professions. This "new face of professionalism" has six features according to Dumont (1970). The first is consumer control, exemplified by the "maximum feasible participation of the poor" characteristic of government Community Action Programs. This feature extends the notion of using client impact as a measure of practitioner effectiveness and makes the client a collaborator or even a controller of the professional. The second feature is an indifference to credentialism often encouraging self-help and paraprofessional programs and the use of indigenous personnel. Carl Rogers, writing particularly about clinical and social psychologists, contends "there are as many certified charlatans and exploiters of people as there are uncertified." He notes that an inevitable result of certification is to "freeze the profession in a past image" since the past decade or two are the source of the content of examining procedures (Rogers, 1973, p. 382). The third feature listed by Dumont is a sense of superordinant purpose in the well-being of people. This is a humanistic concern that causes people to seek out those in other disciplines who share their purposes and to distrust the image of the "expert" who solves problems by virtue of his purely technological skills. The fourth feature--already apparent--is an attitude of criticism and the fifth is impatience with the rate of change and a feeling that social change is literally a matter of survival. Finally, Dumont contends that the driving force for the new professionals is neither money nor prestige nor fascination with "expertism" but compassion.

In a series of essays by "new professionals" most of these characteristics are independently mentioned. There is also the overarching concern with mechanisms for social change (Gross and Osterman, 1972). The editors of that collection take note of the counterreaction to the new professionals, primarily a concern that the new professionals sometimes appear to deny the value of advanced training and highly specialized skills. To that I add the observation that while new professionals have been vocal, they have had little effect. Indeed the movement may have merely been a temporary correlate of the social upheavals during the last decade. Lack of effectiveness may be due to their small numbers or to a tendency, apparent in Dumont's writings, to romanticize their concerns. As we shall see, certification procedures have not yet accommodated the critiques of the new professionals. Interestingly, the New York incident described above suggests that the arena for change may be the legal one where licensing decisions are made rather than the professional one where certification decisions are made.

In conclusion, this report is primarily concerned with how registration and certification agencies may determine readiness for practice in aspiring professionals. We are concerned for the most part with professions in which the practitioner interacts directly with the ultimate client. We are concerned further with interpersonal helping interactions, that is with the helping professions. Thus less attention is given to the professions of journalism, architecture, engineering, and business than to medicine, nursing, clinical psychology, teaching, social work, ministry, and law.

II. Identifying and Selecting Criteria

In assessing individual readiness for professional practice, we are seeking to identify persons most likely to be successful practitioners. What are the earmarks of successful practice? What criteria comprise a definition of professional effectiveness? This chapter reviews procedures by which such criteria are identified and selected. Subsequent chapters describe procedures by which criteria are operationalized for data collection.

Three categories of criteria are customarily identified (for example, see Thorndike, 1949). They are the immediate, the intermediate, and the ultimate. A dental school admissions committee, for example, wishes to admit students who will become effective dentists (ultimate). But because of the complexity of obtaining measures of that criterion, the committee usually assesses its admissions procedures against completion or non-completion of the training program (intermediate) or merely against first year or first semester grades in the program (immediate). Thus distinctions among these criterion categories are a) time specific, as implied by the categories' names, and b) data specific, since the data collected for each category differ in kind.

The assessment of readiness for practice focuses on the termination of training rather than on its beginning. Therefore, our concern in the above example is solely with what was called there the ultimate criterion. The temporal dimension still applies, of course, and we might redefine the ultimate criterion as career-long effectiveness, the intermediate as effectiveness after ten years, and the immediate as effectiveness after one year. Note, however, that across criterion types the data will be less diverse than in the dental school admissions example.

In both cases, criterion data are gathered subsequent to decision data. That is, what is predicted properly occurs subsequent to assessment. Yet the most frequent method of determining the adequacy of tests used for certification and registration decisions is to compare scores on the tests with data gathered earlier in time, during training. I find no psychometric justification for that procedure, if only because it fails the test of relevance. For the immediate criterion to be relevant "it must adequately represent important aspects of the ultimate criterion" (Super and Crites, 1962, p. 34), and performances from which grades are derived are unlikely to be representative of the activities of professionals.

There are many practical reasons that ultimate criteria are so little used. "Even when it is possible to define such criteria precisely enough to obtain measures, the collection of such measures is so time-consuming and expensive as to be practically prohibitive" (Wiggins, 1973, p. 40). Rather than castigate those who have not used such criteria, our purpose is

to praise those who have used them and to point out compelling reasons for continuing the endeavor. Therefore, this chapter is concerned only with criterion situations which occur subsequent to certification.

Techniques which identify criteria by documenting a) what professionals actually do, and b) what duties are perceived as important are here classified into four categories: surveys, observations, correlational research, and idiosyncratic suggestions. Illustrations of each are presented below but no claim is made that these citations are exhaustive.

Surveys

Survey data may be collected from a great variety of sources. The college of peers (Knox, 1973) is a group of highly functioning professionals in the same field. Once these persons are identified they are surveyed regarding functions, duties, etc. Or data may be sought from supervisors, clients, the public at large, and institutional records.

Fifty pediatric cardiologists, stratified by setting, met in five workshops of ten each (Adams et al., 1972). The task for each of the workshops was respectively 1) to define general areas of knowledge and skill necessary for the practice of pediatric cardiology, 2) to rank general areas according to importance, 3) to identify specific components within general areas, 4) to operationally define components, and 5) to designate competency levels for each general area. Each of the first four but not the fifth workshop met its goals.

The Academy of Parish Clergy "has developed a preliminary catalog of competences covering the many arts a parish clergyman should possess" (Adams, 1971, p. 106). Items originally suggested by Academy members were refined in discussions at meetings and in interviews (Thomas, undated).

Hale (undated) surveyed 380 members of the Lutheran Church in America including theological professors, synodical presidents, laity, and clergy peers. The two-page mail questionnaire included the questions "What does 'readiness for ministry' mean to you?" and "Describe an incident that describes this." Results with emphasis on the definitions rather than the incidents were organized into a tentative taxonomy of traits, qualities, and/or competencies.

In a study assessing patient care "a large number of practicing physicians were asked what they considered to be the basic factors of success in their field or specialty" (Price et al., 1971, p. 230). The list of responses for 372 doctors was then submitted to "over 100 selected individuals--medical educators, college and medical students, administrators, clergymen, a variety of other professional people, and patients recuperating in or recently discharged from hospitals" (p. 230). The final edited list of 116 qualities including both positive and negative attributes was formed into a rating scale found to be reliable for rating physicians by groups such as those listed above. Rank orders of these qualities for all respondents (N=1604) and subgroup comparisons have been reported.

In some professions institutional records are an appropriate source for data on typical functions. At a community hospital researchers derived a list from such records of the ten most frequent diseases. They had hospital staff identify criteria for treatment of these diseases and then assess patient charts according to the criteria. Regarding the use of antibiotics, for example, a review of fifty consecutive cases revealed that antibiotics were used correctly in only thirty percent of the incidences. This finding was used to develop programs of continuing education for the staff (Brown and Uhl, 1970).

The health fields have pioneered the use of these procedures, calling them the "health care audit" (Knox, 1973, pp. 37-40). They could, however, be adapted for assessment purposes in any profession which has internships, where careful records are kept, and where criteria of desirable performance may be specified.

Another type of survey is the critical incident technique. Since its description by Flanagan (1954), the critical incident technique has been widely used in studies in education and other professions. The technique asks for explicit and objective descriptions of events. An incident is critical if it led to a clearly desirable or undesirable consequence.

Thomas (undated) used group interviews with Washington, D. C. ministers. They selected the four most important competencies from the list derived by the Academy of Parish Clergy mentioned above, ranked them, and then were asked to describe positive or negative instances from their professional experience. Although instances were described less specifically than desired, a list of items was derived from them for a "Pastor Behavior Rating Scale."

The study of certification procedures in orthopedic surgery (Blum and Fitzpatrick, 1965) provides a large scale use of the technique. Incidents were collected from several hundred orthopedists in a variety of settings primarily by a mail questionnaire. Four incidents, two effective and two ineffective, were solicited, each on a separate page. Instructions were detailed providing space to describe a) backgrounds to the incident, b) characteristics of the physician involved, c) exactly what the physician did, d) why the action was especially effective (ineffective). The resulting 1761 incidents were sorted through revisions into 94 categories.

To identify distinctive characteristics of a particular profession compared with other professions, a useful discussion starter is the question "How are clinical psychologists (say) unique?" (Koewing, 1972). This process may identify areas for futher data gathering or for comparison with data gathered by the critical incident technique.

Observation

Practitioners may contribute information on their activities by self-report or through the reports of trained observers. A multiple-choice

questionnaire was devised to collect data on clergymen's distribution of time. Results served as a baseline against which to evaluate the impact of "action training" programs for practicing clergy (Winter and Mills, 1971).

Less subject to errors of selective recall are activity diaries. In pediatric cardiology (Adams, et al., 1972), 192 of the 207 Board-certified pediatric cardiologists kept diaries for blocks of five consecutive professional days. Entries formed clusters around five major areas: professional, research, clinical, educational, administrative, plus a miscellaneous category.

In the same pediatric cardiology study, fourth year medical students were trained in observation skills by a special film. A stratified random sample of 16 of the 207 Board-certified pediatric cardiologists were observed for two workdays and their activities classified according to the five categories derived from the diaries described above.

Week-long observations were made of 20 randomly selected internists and general practitioners in both group and solo practice in San Diego, California. Four medical students made the observations using pretested data recording forms and procedures. Observers were rotated among subgroups of physicians, even recording the exact amount of time that they spent in conversation with the physicians (Brody and Stokes, 1970).

Correlational Research

Several scholars have devoted their energies to reviewing research on teacher performance in order to identify notable findings and empirical generalizations. One approach is to identify teacher behaviors which have been shown to correlate with student achievement. That is, the researcher identifies intermediate criteria which, to the extent that they correlate with student achievement, serve as proxies for the latter ultimate criterion. Another useful distinction made in educational research is that between process and product research, where observed teacher behaviors are process variables whose correlations with student achievement (the educational product) may be investigated. Rosenshine's summary shows the nature of these correlational findings.

> Of all the variables which have been investigated in process-product studies to date, five variables have strong support from correlational studies and six variables have less support but appear to deserve future study. The five variables which yielded the strongest relationships with measures of student achievement are: clarity, variability, enthusiasm, task orientation and/or businesslike behavior, and student opportunity to learn. The six less strong variables are: use of student ideas and/or teacher indirectness, use of criticism, use of structuring comments, use of multiple levels of discourse, probing, and perceived difficulty of the

course. The relationships are positive for ten of the variables and negative for use of criticism.

[By way of contrast he also lists] the behavioral characteristics, equally virtuous and "obvious," which have not shown significant or consistent relationships with achievement to date. These variables...are listed below, and the method by which they were assessed follows in parenthesis: nonverbal approval (counting), praise (counting), warmth (rating), ratio of all indirect behaviors to all direct teacher behaviors, or the i/d ratio (counting), flexibility (counting), questions or interchanges classified into two types (counting), teacher talk (counting), student talk (counting), student participation (rating), number of teacher-student interactions (counting), student absence, teacher absence, teacher time spent on class participation (rating), teacher experience, and teacher knowledge of subject area (Rosenshine, 1971, pp. 54-55).

To what extent is this approach useful for investigating readiness for professional practice? It appears highly appropriate until we acknowledge several important qualifications. First, such results are available for few fields, primarily for teaching and counselor education. In fact, schools of education may be the only professional schools where pressure to publish and the conventions of statistical analysis combine to yield correlational study after correlational study. Second, even in teacher education these reviews are limited to a single ultimate or product criterion, student achievement. Few agree that student achievement alone is a sufficient condition by which a teacher or a school is to be evaluated. Third, reported correlations are less than unity so that the intermediate criteria are not complete proxies for the ultimate criteria. Finally, even if correlational results were high, these studies are no substitute for controlled experimental research. A methodologically proper study would train a group of teachers to exhibit warmth, for example, observational methods would verify that this behavior occurs at a higher level among trained teachers than among controls, and relevant outcome measures would be collected (Rosenshine, 1971).

These problems seriously limit the present usefulness of correlational studies for assessing readiness. Nevertheless, these studies do represent a considerably greater data base than is true for the other techniques described above.

Idiosyncratic Suggestions

A glance through the literature quickly reveals a number of notions which one or more persons contend are significant characteristics of successful professionals. For example, the "fully-functioning person" and the "self-actualized person." These suggestions are usually based loosely

upon the writer's theory or value orientation and may not be rigorously researched. They often reflect unconventional views of the writer and may be of considerable heuristic value.

Of the two I shall describe here, the first sees teachers as models. "Through personal example, and customarily without fanfare, they make the idea of mastery credible and its attainment close at hand. They are, or should be, useful to students not simply as storehouses of information or as skillful performers (for books and films will do those jobs as well) but as model knowers and doers whose physical presence across the room brings the knowable and do-able almost within reach"(Jackson, 1971, p. 24). This is an appealing, albeit romantic notion for any profession.

A second example also drawn from teacher education, sees the effective professional as the "intentional" individual. The person who acts with intentionality "possesses the ability to act on his environment. He can generate alternative behaviors in a given situation and approach the problem from different vantage points as he receives environmental feedback. The intentional individual is not bound to one course of action but responds aptly to ever-changing life situations" (Ivey and Rollin, 1972, p. 161). As a performance criterion intentionality may be relevant and researchable.

Such idiosyncratic notions found in the literature may be suggestive for postulating criteria of successful professional practice.

* * *

Data gathered by any of the above techniques may be judged useful, relevant, and statistically reliable. But an additional characteristic must also be considered: validity. Are the functions generated by the critical incident technique or by diaries or from Jackson's "model" valid? For whom? Under what conditions? That is, do they represent crucial activities of the professional?

Some aspects of this question are relatively simple. The representativeness of a sample of persons surveyed is easily estimated, and likewise the representativeness of settings surveyed. But how can the Zeitgeist be transcended? What is valued by the expert may be denegrated by the client. What both agree upon today may be in disagreement with the preferences of their counterparts in another time and/or in another place. Even within contemporary culture disagreements among experts are frequent. "... the Far West Regional Laboratory has a training program in Flanders' Interaction Analysis which lists more teacher repetition of student answers as one measure of the preferred 'indirect teaching.' How can universally rather than relatively valid criteria be identified?" (Rosenshine, 1971, p. 40).

Throughout history philosophers and religious leaders have argued that particular human states or virtues are universal. Modern philosopher-psychologists are no exception. We shall refer briefly to three developmental stage schemes to which the authors claim that all humankind is subject.

Jean Piaget's seminal work on cognitive development postulates four stages which are said to occur in an unvarying sequence, each of which is at a higher or more desirable level, and which hold across cultures (Flavell, 1963). They are the sensory-motor, preoperational, concrete operations, and formal operations, the last usually reached in adolescence. The epistemological basis of the scheme allows him to go beyond description to prescription, thus claiming what ought to be. Educators, especially in America, have attempted to accelerate development using Piagetian problems for evaluation.

David Hunt and others have originated conceptual systems theory (Harvey, Hunt, and Schroeder, 1961) and the conceptual level approach, a subsequent modification (Hunt, 1971). The scheme is presented as a developmental one only tentatively, since longitudinal data are lacking. Further, it applies only to conceptual organization, a primarily cognitive characteristic, rather than to the developmental stage of the "whole" person. The assumption of each successive stage being more desirable is, however, clearly implied. To determine a person's conceptual level (CL) a paragraph completion exercise is used. The subject gives a personal reaction of three or four sentences to six topics, including "What I think about rules...," "When someone disagrees with me...," "When I am told what to do..." A scoring manual gives instructions for assigning a given answer to one of four levels:

> For a score of 0: very undifferentiated response, overgeneralized exclusion of any negative input, lack of affective control.
>
> For a score of 1: categorical judgments, overgeneralized and unqualified acceptance of single rule, recourse to external standards.
>
> For a score of 2: some form of conditional evaluation, beginning self-delineation, expression of alternatives.
>
> For a score of 3: taking two viewpoints into account simultaneously, coordination of evaluation of situation with differential response, and clear indications of self-delineation and internal standards (Hunt, 1971, p. 37).

Since correlations between these scores and measures of age, class, sex, intelligence, academic aptitude and personality are low (usually under .30), CL is considered to measure a unique construct and has been used in numerous training situations, always valuing the higher level (Hunt, 1970, 1971).

Finally Kohlberg has described a scheme specifically related to moral development in which he claims there are six stages which, as documented in his research, occur in every culture (Kohlberg and Turiel, 1971; Kohlberg, 1973).

> ...the developmental progression that emerged from Kohlberg's analysis proceeds from an initial preconventional level through a conventional level to a principled level. In moving through the stages, the child's moral judgments become increasingly differentiated and integrated. At the first stage, morality

> is defined on the basis of the consequences of actions; moral value is undifferentiated from the value of material objects or from the power to implement rules...At stage 2 there is differentiation of power, material value, etc., from the needs and wants of the self (and to a limited extent other individuals), but moral value is undifferentiated from individual desire or need. At stage 3 a consideration of the welfare of others and conventional role-expectations begin to appear. There is a differentiation of needs and desires from a globally defined "good person" role (based on the approval and disapproval which stem from adherence to--or violation of--conventional expectations). At stage 4 moral judgment is articulated into a concern for, and an understanding of, rules, authority and the social organizations that govern human relations. However, at this stage commitment to the institutional order (a commitment to custom and law for their own sake) is undifferentiated from the principles of universal rights, which can be defined independently of any given social order. At the most advanced level, that of principled morality, there emerges a differentiation of universal moral principles from the rules and conventions of a particular social order. At this level (stages 5 and 6), moral judgment is characterized by an understanding of (and commitment to) values having universal, prescriptive applicability (principles that are not limited to any particular social order), and a differentiation between these principles and values that are specific to a given social order. Therefore, change to this highest level includes a recognition of the arbitrary nature of conventional values in conjunction with greater understanding of the universality of moral principles (Turiel, 1973, pp. 733-734).

These stages are presented as "universal patterns or principles of moral thinking that progress through an invariant order" (Kohlberg, 1973, p. 8). Because stages are typically assessed through interviews in which the subject answers questions about stories of moral conflict, the scheme is cognitively loaded. Moral judgment, not moral action, is assessed. Of course, the former is necessary for the latter.

Because these attempts to document universal sequential stages of development are still in early formulations, they are discussed here primarily because of their heuristic value. If, for professionals in specified settings, particular stages can be identified as essential for effective functioning, schemes such as these may provide means for assessment and perhaps for training. To the extent that the stages <u>are</u> universal, there is reduced likelihood that criteria would be determined by subjective, culturally relative values. (Of course, the selection of minimal stages itself remains subject to parochial bias.)

III. Assessing Personality Characteristics

The use of paper-and-pencil instruments and situational tests to assess personality constructs has had a long psychometric history. The most ambitious personnel selection programs have focused on assessment of personality. Such investigations have consumed large sums of money and enlisted many of the finest social science researchers. Four of them are summarized here: The Office of Strategic Services in World War II, the Institute of Personality Assessment and Research, the Veterans Administration Study of Clinical Psychologists, and the Peace Corps. Although each investigation includes some non-personality variables, the terminology used, especially in describing criteria, places each study well within the personality assessment tradition.

The major source for these summaries is Chapter Eleven in Wiggins' recent Personality and Prediction: Principles of Personality Assessment, although some primary sources are cited as well.

Office of Strategic Services

General task. The OSS was charged with the responsibility of selecting from a total of 5,391 candidates those who would be successful in one of the variety of positions in the special strategic forces of the Armed Services. Persons were measured in groups of 18 over a three-day period, with such factors as the following receiving high priority: success or failure of a mission often depended on the individual, security leaks were intolerable, impaired relationships with resistance forces could be devastating, and morale among OSS units was very important. Therefore, very careful selection with as few "false positive" decisions as possible had to be insured (Office of Strategic Services, 1948).

Criterion analysis. Initially, OSS tried to base their selection criteria on each particular position to be filled, but poor communication from the front lines made this impractical. Interviews with branch chiefs and administrative officers in Washington resulted in the following set of "general qualifications."

General Qualifications

1. *Motivation for assignment:* war morale, interest in proposed job.
2. *Energy and initiative:* activity level, zest, effort, initiative.
3. *Effective intelligence:* ability to select strategic goals and the most efficient means of attaining them; quick practical thought—resourcefulness, originality, good judgment—in dealing with things, people, or ideas.
4. *Emotional stability:* ability to govern disturbing emotions, steadiness and endurance under pressure, snafu tolerance, freedom from neurotic tendencies.
5. *Social relations:* ability to get along well with other people, good will, team play, tact, freedom from disturbing prejudices, freedom from annoying traits.
6. *Leadership:* social initiative, ability to evoke cooperation, organizing and administering ability, acceptance of responsibility.
7. *Security:* ability to keep secrets; caution, discretion, ability to bluff and to mislead

Special Qualifications

8. *Physical ability*: agility, daring, ruggedness, stamina.
9. *Observation and reporting*: ability to observe and to remember accurately significant facts and their relations, to evaluate information, to report succinctly.
10. *Propaganda skills*: ability to apperceive the psychological vulnerabilities of the enemy; to devise subversive techniques of one sort or another; to speak, write, or draw persuasively.

(Reproduced from Wiggins, 1973, p. 523)

Models and strategy. The three-day assessment period included a variety of assessment procedures which allowed for overlap of particular variables so that multiple measures were possible. The assessment personnel consisted of a senior staff of seven psychologists, psychiatrists, and sociologists and a junior staff of eight graduate students in psychology. Three-man assessment teams were assigned to a six-man group of candidates for interviewing, observation, and rating, which was reported back to the entire staff for final decision-making. No particular personality theory was espoused by the staff, but a "holistic" approach to the entire personality of the candidate was used, rather than a more mathematical or specific approach such as the "multiform organismic system of assessment" or an "atomistic" approach.

Data collection. The schedule of testing procedures is found on the following page. Of particular interest is the use of psychodrama with situations tailor-made for each candidate based upon perceived weaknesses in his personality or performance. These situations were designed to elicit samples of behavior considered relevant to eventual criterion performance in OSS assignments. Each three-man staff team was responsible for the data collected for each candidate, but the opinions of all the staff who had contact with the candidate were considered in the final decision.

Data analysis. Because no information was available concerning the empirical validities of assessment procedures, statistical techniques were not used for data combination. Grounded in the "holistic" approach in which the significance of any single assessment datum must be evaluated in relationship to all available data, as well as in relationship to the anticipated criterion situation, the staff met together to consider the team's report and to combine opinions. A large chart displayed all the ratings, based on 35 different data collection procedures, in such a way that both the central tendency and variability of each rating was apparent from inspection.

The decisions made by the staff included the following: accept or reject a candidate, rejection for a particular assignment but accepted for a different type of work, a "red flag" meaning that a candidate should be accepted only on the condition that he perform well in OSS training,

ORIENTATION PERIOD

Briefing in Washington, D.C.
Arrival and orientation at Station S

Dinner

Otis Test of Mental Ability, sentence completion test, health questionnaire, work conditions survey, vocabulary test, personal history form
Belongings test

FIRST DAY

Breakfast

Instructions for terrain test
The Brook
The Wall
Construction; postconstruction interview
Personal history interview (X condition)

Lunch

Propaganda skills, map memory test, Bennett Mechanical Comprehension Test, Instructions for Manchuria Propaganda Assignment
Leaderless group discussion; sociometric assessment

Dinner

Instructions for interrogation test
Stress interview
Poststress interview
Test of ability to evaluate and to abstract intelligence information
Preparation for teaching test

SECOND DAY

Breakfast

Assigned leadership tasks: Mined Road, Sentry, Agent's Room, Blown Bridge, Killing the Mayor
Teaching test
Terrain test

Lunch

Names and identifications test, movie observation test, code aptitude test
Interrogation test
Obstacle course

Dinner

Improvisations

Cocktails

Debate

THIRD DAY

Breakfast

Sociometric questionnaire
Personality sketches
Instructions for murder mystery test

Lunch

Athletic events
Baseball game
Murder mystery court

Dinner

(Reproduced from Wiggins, 1973, pp. 528-9)

and detailed ratings of the candidate's fitness for different conditions of assignment, different levels of responsibility, and different types of assignment.

Outcome of prediction. Four methods of obtaining measures of overseas performance were used: an overseas staff appraisal, a theater commander appraisal, a reassignment area appraisal at a relocation center after one tour of duty where an interview and self-report was given, and a returnee appraisal by the assessment staff in Washington in which the returnee was asked to rate the performance of persons with whom he had been acquainted. Intercorrelations among the measures ranged from .49 to .59. Overseas staff appraisal, considered the best performance measure, had a .37 correlation with the assessed job rating with 15% of those assigned considered unsatisfactory (N=88). Basic problems lie in the fact that there is no baseline from a previous selection procedure to use in comparison with these figures, that only 4 to 7% of those appraised in performance were screened after the full assessment procedure was put into effect, and that the appraisal procedures were not fully developed until the war was nearly over.

Wiggins (1973) takes the report of the OSS assessment staff and by a mathematical procedure based on the .75 selection ratio, using the .37 correlation as an estimate of the phi coefficient, and using the formula for estimating a base-rate, he concludes that 77% of the decisions made during the three-day assessments were correct, which is an increment of 14% over random selection. False positive decisions were reduced by 7% over random selection.

A similar assessment procedure also used by OSS but requiring only a one-day testing period using similar but abridged methods was more successful in predicting success in OSS assignments. Wiggins computes its overall proportion of correct decisions at 84% with a false positive rate of only 4% (1973, p. 538).

Critique and significance. A major value of this study is its self-critical nature which has made it a reservoir of suggestions for improvement and further development since 1948.

An interesting speculation of the OSS report writers was that perhaps the reason why the three-day assessment period was less effective than the one-day period was that the three-day period provided too much information about the candidates relative to the available information about the job descriptions of the assignments. The "excess" information about the candidates may have produced confusion. If this is true, the necessity of balancing the amount of information needed to make the assessment relative to the job criteria may be an important factor.

Wiggins (1973) emphasizes that simple correlations may be very misleading, and his formula for estimating the base rate and making percentage comparisons with it is a valuable estimate providing his assumptions are correct.

Institute of Personality Assessment and Research

The Institute of Personality Assessment and Research was developed by several of the staff of the Office of Strategic Services after finishing their service in the military and is housed on the University of California, Berkeley, campus. One of their many personality assessment studies is described here.

General task. Two basic questions determined the scope of this project: "What patterns of ability, motivation, and interest are associated with outstanding effectiveness of Air Force officers in their military assignments?" and "By what psychological tests and procedures or by what combinations of such techniques can the distinguishing characteristics and the potential promise of the effective officer be identified and measured?" (MacKinnon, 1958, p. 1).

Criterion analysis. Sources of criterion measures were promotion board ratings, officer effectiveness reports, superior officers' ratings, and job concept interview ratings. These 61 different criterion ratings were reduced to eight on the basis of factor- and cluster-analytic procedures. No single factor of "general effectiveness" obtained, but three distinct factors did emerge: 1) General effectiveness as evaluated by commanding officers; 2) General effectiveness as rated by trained psychologists; and 3) Task accomplishment at the expense of interpersonal acceptance (as evaluated by immediate superiors). The first of these became the pivotal reference variable. A criterion index was defined as the algebraic sum of scores on these three dimensions.

Models and strategies. The general strategy of this study is characterized as an empirical one, although the usual procedure used by IPAR of selecting contrasting high- and low-performance groups from which to determine measurable differences was not employed. However, the 100 officers (from the 343 who were field-tested) participating in the living-in assessment were divided into the 25 highest-scoring and the 25 lowest-scoring on each variable and were compared on other variables.

The basic model of the "ideal Air Force officer" was developed from a "personological job analysis" and from previous assessment research programs. The ten components considered essential are listed on the next page.

Data collection. Data about the Air Force captains were collected in three stages. The 343 men tested at seven local bases were given a battery of 27 different paper-and-pencil tests which were measures of 1) biographical variables, 2) personality and interests, 3) cognitive and intellectual functions, and 4) social insight and judgment (e.g. SIVB, CPI, and MMPI) from which 233 separate scores for each officer were obtained. The second stage involved a three-day living-in assessment in which additional paper-and-pencil tests were given as well as projective measures such as the Rorschach and the TAT. Detailed life-history interviews, physical examinations, and perceptual performance tests (e.g. rod-and-frame test) were also administered. The third stage of assessment involved ratings by

1. *Soundness as a person.* Maturity in personal relations; self-insight and self-acceptance, as well as acceptance and understanding of others. Absence of serious emotional problems. Stability of mood and manner. Good balance of social conformity and spontaneity.

2. *Intellectual competence.* Effective intelligence. The ability to perceive and to solve problems, with particular emphasis on implications for action. This involves, first of all, clarity and intellectual power, but a factor of practicality and translation of cognition into action is also important.

3. *Good judgment.* Soundness and good judgment in evaluating self and others. Wisdom, the ability to see group situations and individuals in broad perspective and to draw dependable and practical inferences; capacity for independent thinking, and a willingness to make decisions on one's own.

4. *Health and vitality.* Consistent good health and stability of physiological functioning. Absence of minor as well as major illnesses. Resistance to disease. Ability to withstand stress and endure hardship, privation, and insalubrity. Vigor, robustness, stamina, sense of physical well-being.

5. *Military and social presence.* Poise and self-assurance in dealing with others. Impressive, able to command both attention and respect from observers. Executive manner—a matter-of-fact attitude of expecting one's orders to be obeyed and carried out. Decisiveness, absence of confusion, self-acceptance without egotism and self-confidence without arrogance.

6. *Personal courage.* Ability to meet danger without undue fear. Resourcefulness under stress. Absence of any tendency to side-step troublesome situations or to make concessions merely to avoid conflict. An appetite for hazard and risk-taking, but without foolhardiness or rashness. Willingness to commit oneself in a possibly dangerous situation.

7. *Originality.* Originality and creativity of thinking and in approaches to practical problems; constructive ingenuity; ability to set aside established conventions and procedures when appropriate; a flair for devising effective and economical solutions for perplexing problems; the knack of capitalizing on the odds instead of being dominated by them.

8. *Fair-mindedness.* Candor, forthrightness, honesty, impartiality, objectivity and sympathy. Ability to judge issues without bias, enmity, or spite, and to avoid dogmatism and prejudgments.

9. *Integrity and responsibility as a commander.* A linear relationship between inner ethicality and outer manner. Absence of subterfuge and deceitfulness. Effortless acceptance of superego values, and ability to project them as sensible and worthwhile. True respect for others; unembarrassed acceptance of personal dignity in self and others.

10. *Positive valuation of the military identity.* Capacity for loyalty and for devotion to military and patriotic ideals. Conviction about the worth of military activity and function; ability to tolerate frustration and discouragement and to maintain perspective and goal orientation under stress.

(Reproduced from Wiggins, 1973, pp. 544-5)

trained observers and psychologists, such as group and situational tests of decision-making behavior, charades, improvisations, and group pressure for conformity of judgment. Thirty personality variables, a 76-item Q-sort, and the 300-item Adjective Check List were used by the six to eight psychologists who observed each officer.

Data analysis. Stanine scores were used to reduce computational difficulty, but more than 5,000 Pearson product-moment correlation coefficients between predictor and criterion variables were still required. To insure that a significant variable (.05) was not simply a result of the chance factor in such a large number of computations, an internal check was devised in which a variable was not considered significant unless it appeared in all three random subgroups of the original 343 officers, and in both random subgroups of the 100 living-in officers.

Outcome of prediction. The results of this study are disappointing overall. "Strictly speaking, the predictor-criterion relationships were postdictions, because most of the criterion indices were obtained well before the time of assessment, in some instances as much as five years earlier" (Wiggins, 1973, p. 547).

When the field-testing variables were correlated with the seven components of officer effectiveness reports (job knowledge, cooperation, judgment, responsibility, leadership, growth potential, overall), six percent of the 4,000 correlations were significant, but none of them obtained in all three subgroups. "Work effectiveness and responsibility," one dimension of job concept interview ratings, was predictable from 52 of the 233 field-testing variables. The criterion index was also predictable from 23 percent of the field-testing variables, although none were significantly correlated in all three subgroups. A cluster analysis of the 21 most promising predictors of the criterion index revealed four cluster dimensions which when correlated with the criterion index had a correlation of .32 in a sample of 100 officers. IPAR conclusions were that 1) the most promising predictor scales for officer effectiveness came from empirical personality and interest inventories, 2) the field-testing composites derived from a conceptual model had the highest "survival rate" of any single source of data, 3) a battery of 15 tests taking approximately 10 hours to complete would include all of the most promising predictors in the field-testing variables.

In the living-in assessment, none of the variables were significantly correlated with effectiveness in both subgroups. A combined significance of .05 was obtained when five of the 299 variables were correlated with general effectiveness as reported in superior officer's ratings: 1) percentage of whole responses on the Rorschach, 2) guessing effectiveness in a charades task, 3) age of subject, 4) pathogenicity of childhood, and 5) staff Q-sort on "undercontrolled." Again "work effectiveness and responsibility" from the job concept interview ratings was predictable (99 of 398 living-in variables). The criterion index was predicted by 51 of these variables, but was considered barely significant in both subgroups. From

a clustering of the 20 most promising predictors into four dimensions, a correlation of .47 was found with the criterion index in a sample of 100 officers. "The most promising living-in assessment variables were the pooled staff ratings on 30 dimensions and 76 Q-sort items," and the situational tests and perceptual tests were considered promising. "In general, it was concluded that a living-in assessment of at least one day and preferably two to three days would yield valuable data for the forecasting of officer effectiveness" if the expense was deemed worthwhile (Wiggins, 1973, p. 550).

Critique and significance. This IPAR study is not an empirical strategy study in its purest form because contrasting groups were not used, and thus the development of new instruments and different combinations of predictor variables was not possible. The investigators did use almost every type of personality assessment procedure known to be relevant to a social setting. The criterion index measured generalized "officer effectiveness" rather than specific component skills. With greater specificity perhaps a better understanding could have been reached regarding individual attributes and effectiveness. The major contribution of this study lies in the area of its methodology, since it was not a traditional selection study (prediction of future performance) or a study in which concurrent status on differentiated criteria could be obtained. The promising nature of the staff ratings of the three-day assessment indicates that perhaps a diagnostic council approach would be a highly useful tool in assessment.

Veterans Administration Selection Research Project

General task. The task of this research study was to identify the most promising procedures for the selection of graduate students in clinical psychology for Veterans Administration traineeships. Several procedural questions left unanswered by the previous two studies were also explored in this study: 1) living-in assessment approach versus an empirical test-battery approach, 2) pooled staff ratings versus individual judges, and 3) the relative worth of differing amounts of input information in clinical prediction. The study was conducted in two parts, a "1947" group who had already been selected for graduate training and who participated in a five-day living-in assessment, and a "1948" group who did not know of their acceptance or rejection for graduate training and who did not have a living-in assessment (Kelly and Fiske, 1951; Kelly and Goldberg, 1959).

Criterion analysis. One questionable characteristic of this study is that the criterion development took place after the assessment had been made, and a basically a priori definition of duties was used by the assessment staff. Duties such as the following were included in the ratings: 1) clinical diagnosis, 2) individual psychotherapy, 3) group psychotherapy, 4) research, 5) administration, 6) supervision, 7) professional relations, 8) integrity, and 9) overall suitability for clinical psychology. The criterion measures used constituted one of the most extensive analyses ever employed in the study of a profession and included the following: 1) academic performance ratings, 2) supervisors' ratings, 3) content examinations, 4) work

sample measures, 5) a trainee experience inventory, and 6) a follow-up questionnaire mailed to the subjects. However, the assessment measures were chosen before these criterion analyses were available and without their benefit.

Models and strategies. While both clinical and statistical strategies were to be used in this study, neither was employed to the full extent in that situational analysis, recommended in the clinical setting, and contrasting groups in the empirical approach were not used. No single personality theory was used in developing the model, and several of the judges later became noted representatives of varying views.

Data collection. The format for the five-day living-in assessment for the 128 entering VA trainees in the "1947" group was essentially the same as that for the OSS study above. The list of measures used is found on the next page. Variables for the rating scales used by the staff included a set of 22 phenotypic variables (e.g. "assertive versus submissive") and 10 genotypic variables (e.g. "appropriateness of emotional expression"). Also, 11 criterion skills were rated as a measure of future performance (e.g. "effectiveness in individual psychotherapy").

The "1948" group of 545 applicants for VA traineeships were administered the Miller Analogies, Strong VIB, Guilford-Martin, Allport-Vernon, Sentence Completion, and Rorschach tests. Credentials files and biographical material were collected.

Outcome of prediction. While it is known that these procedures had a 20 percent overall selection ratio for determining the success of the 545 applicants tested in the "1948" group in being accepted for traineeships, other prediction measures were not built into the study. Zero-order correlations were computed between criterion measures and the test scores and staff ratings. All the results were discouragingly low. The Miller Analogies Test was predictive of scores on the content examination in both groups, and in the "1947" group was also predictive of academic performance and supervisory and clinical competence ratings. Overall, the objective tests had higher correlations with their corresponding assessment ratings (e.g. staff ratings of integrity with criterion ratings of integrity). However, the best single predictor was the staff rating of the variable of "individual psychotherapy." Other conclusions were summarized by Kelly and Fiske (1951, pp. 195-196): predictive ratings show fair to high interjudge reliability when based on the same materials. Assessment predictions based on the credential file plus the objective test profile tend to be almost as accurate as those based on more information. Pooled ratings in staff conference are not significantly superior to arithmetical combinations of ratings. Self-ratings by candidates show relationship only to the criteria of intellectual success. Observation by assessment teammates and by staff in situational tests have about equal validity with staff ratings based on the credentials file alone. Projective tests taken individually or as an entire protocol have very little predictive validity for success in clinical psychology.

Credentials file
College transcript
Civil Service Form 57
Correspondence
Letters of recommendation
Intelligence and achievement
Cooperative General Culture Test
Miller Analogies Test
Primary Mental Abilities Tests
Interests
Allport-Vernon Study of Values
Kuder Preference Record
Strong Vocational Interest Blank
Personality
Guilford-Martin Battery
Minnesota Multiphasic Personality Inventory
Projectives
Thematic Apperception Test
Sentence Completion
Rorschach
Bender-Gestalt

Interviews
Initial interview (one hour)
Intensive interview (two hours)
Final interview
Biographical
Biographical inventory
Autobiography
Situational
Leaderless group discussion
Improvisations
Block situation test
Expressive movement test
Party
Sociometric
Ratings of three teammates
Character sketches of teammates

(Reproduced from Wiggins, 1973, p. 560)

The follow-up questionnaire to measure success ten years later showed very low predictor-criterion intercorrelations (Kelly and Goldberg, 1959).

Critique and significance. The major contribution of this study seems to be in its thorough criterion analysis and its additional empirical data in answer to the procedural questions raised by the first two studies (OSS and IPAR). However, the severe handicap of conducting the assessment procedures without the completion of the criterion analysis seems to have greatly limited the usefulness of this massive effort.

Peace Corps

General task. The Division of Selection of the Peace Corps has the responsibility of fulfilling the Congressional committment to provide "men and women of the United States qualified for service abroad and willing to serve" (Peace Corps Act of Congress, September 22, 1961).

Criterion analysis. The initial definition of criterion job performance was provided by the specifications the host country used in describing the position to be filled. These job descriptions could be updated as each tour of duty was evaluated. Jones (1968a, 1969) devised a model for studying overseas effectiveness of Peace Corps volunteers which has been used by the Division of Selection. A rating scale of overseas performance was usually completed by the overseas representative after three months; twelve rating scales were used: job competence, relationships with other volunteers, relationships with host national counterparts, relationship with other nationals, and emotional maturity, as well as an overall evaluation.

Models and strategies. The development of the Peace Corps since its beginning has shown a movement from an OSS-like operation to assess generalized suitability for overseas performance to a civil service-like agency matching specific skills to specific job assignments. Also, training programs have become more specialized to match particular overseas situations. About 17 percent of the applicants are eliminated on the basis of the initial application and language aptitude tests, and 20 percent more are eliminated on the basis of the extensive reference check. Only one of six who apply to Peace Corps are admitted to training. Individual interview is not used as a major screening device, and is not employed at all before training. Once training begins, the advisory selection board, essentially a diagnostic council, becomes the screening agent, and decisions are made halfway through and at the completion of training to "reject," "provide feedback," or "transfer." A "feedback" decision means a recycling of the assessment procedure and reevaluation of the data.

The composition of the advisory selection board represents the strategy of decision-making used: 1) decision making (field selection officer), 2) assessment (field assessment officer, psychiatrist, physician), 3) immediate criteria (project director, training officer, training staff), and 4) intermediate criteria (overseas representative, returned Peace Corps volunteers). The board is advisory in that the final decision is made solely by the field

selection officer after receiving all the available information, including a comprehensive civil service investigation of the candidate the report of which only he (the selection officer) sees. The use of returned Peace Corps Volunteers on the selection boards provides for direct communication between the actual assignment situation and the selection procedures for that position.

Data collection. An essential piece of data collected is the five-place ratings (such as language, motivation for Peace Corps, emotional maturity and overall project suitability) from references given by the candidate. These ratings were later found to be promising predictors of overseas performance (Jones, 1969). Within training, as many as 34 different measures of personality and intelligence have been used, and prior to each of the two board meetings, each trainee was typically interviewed at least once by the assessment officer. Trainees were asked to predict the five most successful and five least successful peers in their group, and to indicate the five trainees they would most prefer to work with and the five they would least prefer. The civil service investigation provided an average of 25 pages on each candidate from such sources as police files, credit bureau records, school records, and other sources of information (at a cost of $450 per candidate). Jones (1968b, 1969) pointed out that the full investigation is not essential to predicting overseas performance unless quantified by the use of objective rating scales.

Data analysis. The center of data analysis is the advisory selection board, a unique body in such assessment studies as it centers the decision-making responsibility on one person (field selection officer) and it includes input from both immediate criteria (training personnel) and intermediate criteria (overseas project personnel). Each candidate was rated on a five-place scale for the components of skills, motivation, interpersonal relations, language ability, and emotional maturity. Judgments of the selection officer correlated very highly with the consensus of the other board members. When Goldberg (1966) compared ratings by board members prior to receiving the assessment summaries, after receiving assessment summaries, after receiving assessment summaries but prior to discussion, and after discussion, he found that judgments by the various judges increased at each stage (.66 to .82), that ratings became more critical after discussion, and that the range and standard deviation of the ratings increased (judges became more discriminating).

Outcome of prediction. One measure of the boards' success in predicting overseas performance is the number of volunteers who return before completing their assignment. This figure is somewhere between 11 and 15 percent. Less than one percent returned for psychiatric reasons, .6 percent for medical reasons, and about half for "compassionate" reasons such as family responsibilities, meaning that selection errors were only about three to four percent.

The best predictor of overseas performance was the final board rating made by the field selection officer (perhaps partly because he tends to "weight" relevant technical skills, peer ratings, and the judgments of his

professional colleagues more heavily than the many other sources of information). Peer ratings of overall success were also predictive, as were instructors' ratings. In spite of these encouraging results, it must be understood that a contamination factor lies in the fact that the overseas representative is involved in the board selection, and he also is the criterion rater of field performance. Also, the results of the board deliberations are known to the field assessment personnel through overseas placement reports. The restriction of range in candidates due to the very selective screening process, the coarseness of measurement, and unreliability of both predictors and criteria must also be considered.

Critique and significance. "Peace Corps selection procedures were not exemplary of every principle of personality assessment...but the procedures were surely more sophisticated and more likely to be successful than those of any previous large-scale assessment project" (Wiggins, 1973, p. 600). The rigorous selection ratio indicates a probable large number of false negative decisions, but such a price was deemed necessary to avoid unfavorable publicity and to fulfill the commitment to provide men and women qualified to serve.

Discussion

Results of these four major personality assessment projects offer little encouragement for the use of personality assessment strategies in assessing readiness for professional practice. A 1967 review of research on personnel selection quotes with approval the recommendation that personality measures not be used for the present as instruments of decision (Guion, 1967). In a subsequent review of personality measurement studies, Fiske and Pearson observe, "While some progress is being made, it is apparent that the gains are small relative to the levels of technical adequacy to which we aspire. Obtained correlations are typically low, especially between distinctly different approaches to the same concept. Even if the instruments are similar, the correlations are still low if the occassions, settings, or instructions differ. With all these factors constant for two methods, the relationships are still well below the internal consistencies of the instruments, particularly when the instruments have different authors possessing their personal views of the labeled variable" (1970, p. 76).

Nevertheless, each year aspiring members of the Bar are assessed for "good moral character" (Mackert, 1970). A committee of the Association of American Medical Colleges is working at developing measures of personality assessment (D'Costa, 1973). And batteries of personality tests are completed by each new generation of students.

From efforts such as Peace Corps research, the major imperative for clinical psychology is that it question the persistent conceptual bias "that an expert judge, given the right conditions and relevant nonquantified information concerning the criterion (which no one has ever discovered for a specific problem in prediction), will somehow demonstrate his adequacy as a predictor of future behavior (Harris, 1973, p. 246). More promising, Harris contends,

are devices such as his own empirically developed rating scales of personality attributes used in his Peace Corps research. The use of fictional autobiographies seems also to have been effective in Peace Corps selection (Ezekiel, 1968; Smith, 1966). In one of the few longitudinal studies of health professionals, the Myers-Briggs Personality Inventory, appears to be of predictive value (Briggs-Myers, 1964).

But because of the generally discouraging results and because of the usual high costs of thorough personality assessment, effective identification of readiness may depend on other strategies such as those described in subsequent chapters.

IV. Assessing Knowledge of Subject Matter

Professions by definition are based on intellectual principles, rather than solely on technique. Consequently, the assessment of subject matter knowledge as one component of readiness is almost universally practiced. The relative ease by which subject matter knowledge can be assessed is also responsible for its pervasiveness. Examinations, usually paper-and-pencil, are employed because they are efficient. In this chapter we are primarily concerned with tests of facts, concepts, and principles. In the next chapter we shall consider assessment of ability to apply subject matter. The distinction is somewhat artificial, since both are cognitive (measure only "knowledge about"), both typically use paper-and-pencil instruments, and results of both are usually combined and reported in one score. The difference is primarily that knowledge of subject matter items likely requires only cognitive skills of recall or recognition.

After a consideration of these examinations, we shall turn to the question of examination validity and to the preferred as well as actual criteria for determining validity.

The use of sophisticated psychometric principles in examination preparation varies considerably from profession to profession. A test for clergy (Briner, 1971) seems rather primitive in this regard. Nursing (National League for Nursing, 1970), dentistry (American Dental Association, 1971), and pharmacy (Greising, 1972) have used careful test construction procedures but have gathered little evidence on validity. Teaching (Quirk, Witten, and Weinberg, 1973), law (Covington, 1972), engineering (Hoerger, 1972), and medicine (Hubbard, 1971) have both used careful construction procedures and also have made attempts to assess validity.

In a discussion of tests produced by the National Board of Dental Examiners, Shafer expresses his confidence that "more thought, more preparation, and more statistical analysis on test items enter into the formulation of these examinations than into 99 percent of all tests given in all dental schools in the United States. Then, too, the Board examinations are compounded so as to eliminate individual bias and regional or geographic variations in techniques, theories, ideas, and semantics; to give every examinee an equal and ample opportunity to demonstrate his proficiency in every area of testing; and very importantly to arrive at a fair, mathematic analysis of the reliability of the examinations with respect to their probability of discriminating accurately between the good student, the average student, the poor student, and the dummy" (Shafer, 1968, p. 188). Examinations are evaluated for adequate internal consistency (reliability), for appropriate item difficulty, and for appropriate item discrimination (between high and low scorers). Since items are based on subject areas in the dental curriculum and are screened by dental school professors, the examination is assumed to have content validity. Shafer does not discuss other forms of validity.

The tests of the National Association of Boards of Pharmacy include two parts. The first, "the written theoretical" includes 120 multiple-choice

questions. (The second part on the "practice of pharmacy" is described in the next chapter). Multiple-choice items are written to represent each of the following taxonomic levels: knowledge, comprehension, application, analysis, synthesis, and evaluation. It remains debatable whether items written for higher levels demand cognitive skills other than recall and recognition; no empirical evidence is available. Another feature of the pharmacy manual is a list of competencies, each illustrated by a multiple-choice item. Three examples follow: "Given the chemical name or formula of a drug, the candidate shall be able to predict the pharmacological action of the drug, or vise versa" (Greising, 1972, p. 11). "Given the Henderson-Hasselbalch equation, a log table, and the proper pK values, the candidate shall be able to find the ratio of salt/acid required to prepare a buffer of any desired pH" (p. 19). "Given the results of a culture and sensitivity test, and patient's history, the candidate shall be able to select the antibiotic of choice" (p. 31).

In order to make more efficient the traditional essay Bar examination, a Multistate Examination was developed by early 1972 for use by states (Eckler and Covington, 1971). It includes 200 multiple-choice items equally divided across five subject areas. Each state may provide supplementary questions, usually essay, on that state's law. States receive scores of the objective items from Educational Testing Service within fifteen days and set their own passing scores. All states administering the exam do so on the same date to protect its security. Covington (1972) notes that since states control the results, validity studies will have to be done at the state rather than at the national level.

Since the early 1950's the National Board of Medical Examiners has used predominantly objective examinations. In addition to work with specialty boards, the Board produces a three-part examination. Parts I and II each have 800 to 1,000 items and test knowledge in the basic sciences and clinical sciences respectively; Part III is a shorter one-day examination measuring clinical competence and is described in the next section. All items are multiple-choice and include three types: the one-best response, the matching, and the multiple true-false. Numerous studies of these examinations have been published and summarized by Hubbard (1971). They meet the usual psychometric criteria of relevance, difficulty, and discrimination. In addition the Board requires reliability coefficients of at least .90 (usually Kr 20) if scores are used for decisions about individuals.

The careful use of principles of test construction has probably significantly improved these examinations during recent years. The 1964 Orthopedic Certification Examination was found to have more than half of its items calling only for recall of information. Only 25 percent were thought by any of the raters to require interpretation of data, application of principles, or evaluation (Miller, McGuire, and Larson, 1965). Present standards do not permit the use of such examinations.

The oral examination has had a persistent history in education, especially in medicine. It is discussed as a measure of subject matter because, despite its purposes, it has not been shown to reliably assess the complex

clinical skills for which it is designed. In a study of more than 2,000 specialty board orals, McGuire (1966) found that questions predominantly called for recall of isolated fragments of information and that candidates rarely cited evidence for their answers. Interjudge agreement in such examinations has been low (Nattress, 1964; Foster, 1969; Hubbard, 1971). The traditional bedside examination used in Medical Boards was discarded in 1963 after a three-year study of 10,000 examinations showed agreement between the two judges to be at only the level of chance (correlation of .25).

"Oral tests tend to be highly subjective rather than objective. In spite of this, the oral examiners tend to be highly confident regarding their judgments" (Nattress, 1964, p. 5). The continued use of orals is planned in pediatric cardiology with changes that are intended to increase reliability. There are to be two 55-minute sessions with three examiners. But there are no resources to evaluate effectiveness (Adams, et al, 1972). The American Board of Professional Psychology which awards diplomas in clinical, counseling, school, and industrial and organizational psychology, has surprisingly eliminated both multiple-choice and essay parts of its examination and retained only the orals. Candidates submit one or more samples of their practice (such as transcripts or tapes of interactions with a client). In addition they may be observed in practice. Five examiners conduct the oral which covers four content areas. I have found no research support for this procedure, and it has been criticized by fellow psychologists (Tempone, 1971).

Because tests of application of subject matter are often part of the exams just described, we shall discuss them before addressing the question of examination validity.

V. Assessing Ability to Apply Subject Matter

The distinction between knowing subject matter and knowing how to apply it is a subtle one. The essence of the latter lies in confronting a new situation, one which never occurred during training, but which can be dealt with by an analysis, synthesis, or evaluation of content learned during training. In effect, the candidate is being asked, "What would you do if..."; and the conditions described are not identical to any he has met before. We shall see below a variety of formats, mostly paper-and-pencil, that such tests can take. It is never known, of course, what the candidate would in fact do in such a situation; all that is known is what he says he would do.

The second major part of the pharmacy licensing examination covers the "Practice of Pharmacy." The competency statements for this examination cover receiving and compounding prescriptions, e.g. "receiving a prescription by telephone, the candidate shall be able to accept it rapidly, detect errors of ommission and commission, and request appropriate clarification" (Greising, 1972, p. 31); detecting errors and ommissions, including errors in dosage forms, improper translation of directions, and ommission of compounding data or technique; patient medical profile; clinical toxology; jurisprudence; etc. Although the intention of this examination is clearly to assess higher cognitive processes, specification of competencies may reduce the likelihood that items do present novel situations.

The Patient Management Problem (PMP) used in medical education is a departure from the traditional multiple-choice question. First described by Williamson (1965) the PMP has been refined by the National Board of Medical Examiners and now comprises a portion of Part III of the National Board Examination. A study of the internship gathered 3,300 critical incidents from 600 physicians and sorted them into nine areas, e.g. physical examination, diagnostic acumen, and physician-patient relationship, each with three or four subdivisions. In a PMP the candidate is first given limited information about a patient, as would be the case in real life.

> He must study the available information, and then he must decide what to do. He may require laboratory studies and diagnostic procedures; he must arrive at decisions about therapy and management. In the test booklet, a list of possible procedures immediately follows the description of the patient. Some of these procedures are correct and mandatory for the proper management of the patient; others are incorrect or contraindicated. The candidate is not told how many procedures or courses of action are considered correct; his task is to select those he judges to be indicated at this point in time. After he has decided upon a course of action, he is instructed to turn to a separate answer booklet where he finds a series of inked blocks, each block numbered to correspond to one of the given choices. He removes the ink for his selected choice with an ordinary pencil

> eraser, and the result of his decision appears under the erasure. He is told that information will appear under the erasure for incorrect as well as correct choices; if he has ordered a diagnostic test, the result of the test will appear under his erasure whether or not the selected test should have been ordered.
>
> As the examinee makes his erasures he gains information from his decisions, from the courses of action he has selected. The situation changes: a new problem evolves and new decisions and actions are called for in the light of new information and altered circumstances. A second series of choices constitutes a second problem; again the candidate reaches his decision and turns to the answer booklet to discover, by erasing the appropriate ink blocks, the consequence of his second series of decisions. He continues in this manner through some four to six problems (sets of given choices of action) following the patient's course for days, weeks, or months, until the patient improves and is discharged from the hospital, or possibly dies (Hubbard, 1971, p. 44).

Because each problem poses many decision points for the candidate, scoring is complex. Each course of action is classified as a) good for the patient, b) harmful for the patient, or c) relatively unimportant for the patient. Scoring penalizes for category (b) choices and rewards for category (a) choices while those in category (c) are ignored. Reliability for PMP's used in Part III is .80 to .85, comparable to the reliability of a section of equal length of Parts I or II. Scores on PMP's correlate .34 and .48 with scores on Part II, which suggest that those portions of the exam do measure different qualities and is approximately the relationship expected between knowledge of medical subject matter and skills of application of that knowledge.

During 1973, ten sets of PMP's are being comparatively tested in both paper-and-pencil and computerized format. Computer presentation increases the control of the sequence in which information (problems as well as options) is presented to the candidate and also records the order in which examinee choices are made (National Board of Medical Examiners, 1973).

The Branching Simulation Exercise (BSE) described by Nattress (1970) differs from the PMP as branching programmed instruction differs from linear programmed instruction. BSE's have been used for continuing education in a number of professions. First, the professional is presented with introductory case material about the client.

> The information which is given about the client throughout the exercise is reported in the way a professional in practice would receive it without interpretation. The examinee is called upon to act on the information available as he would in practice. The examinee may seek assistance with the interpretation of data or findings, or request a consultation with a professional in his own or another discipline,

depending on the construction of the exercise.

Each exercise consists of a number of sections, some of which are not relevant for the client in question. The sections are arranged in scrambled order to minimize the possibility of using the options offered as clues to the expected behavior. In each section, the examinee must indicate his decisions about a series of specific actions, and at each stage he must make a decision about the overall management of the client which will determine the section to which he will be directed next. As the exercise progresses, the examinee is called upon to make decisions based not only on his knowledge, but also on the specific responses of the client to the examinee's decisions (Nattress, 1970, p. 2).

Once a choice has been made, even if incorrect, the examinee cannot change it. Since each examinee may adopt his own strategies for a problem (through branching) the problem becomes unique for him. The opportunity for unique patterns of choice generates a great deal of useful information about the examinee but does not permit ready comparison of responses across examinees. Interjudge agreement for BSE's is satisfactory, .71 to .85 (Sedlacek and Nattress, 1972). Work on the complexity of the scoring has included use of the method of absorbing Markov chains (Hoffman and Nattress, undated).

The computer-based examination technique (CBX) project of the National Board of Medical Examiners is another attempt to assess clinical skills in medicine. The computer is used to simulate the patient and the clinical setting in greater detail than paper-and-pencil PMP's can do or than computerized PMP's have done.

As the examinee sits at the computer console, he is given a booklet which lists all the questions the computer is programmed to answer.

> The examinee must first gather all pertinent information about the patient by asking questions related to the clinical history and physical examination. From the general index he selects the questions he wishes to ask. He then enters the numbers of these questions in the computer terminal. The computer responds immediately. For example, if the physician should ask about the patient's weight, the patient (computer) might reply: "I have lost some weight recently." A second question about weight might be asked with the further response: "I have lost forty pounds in the last year." The physician-examinee proceeds in this manner to ask other questions he considers directly related to the patient's problem. The responses printed out by the computer then provide a record of the clinical history and physical examination.
>
> Because several different clinical skills are being measured by the CBX, the test in this initial prototype

model is such that the examinee's responses in the history, physical examination, differential diagnosis, and laboratory sections are evaluated separately. At the end of each stage in the management of the patient's problem, all information that should have been obtained during that stage is given in summary form. After the examinee has obtained a clinical history, the computer will provide any additional essential information that he will need as a base line for reliable measurement of his competence in handling the next stage of the clinical problem. He then proceeds to find out about the physical findings he thinks would be particularly helpful in the light of what he has learned from the history.

Following a review of the summary of the history, and physical findings printed out by the computer, the examinee then formulates a differential diagnosis. He again refers to the index for a general listing of diagnoses. He selects those he considers highly probable and enters them by number in the computer. Then he is given a list of all diagnoses that should have been included in the differential diagnosis, and is again brought to the same base line and confronted with the same problems in the further assessment of his use of diagnostic studies and procedures.

The next step provides the examinee with a simulated opportunity to use a large hospital laboratory and the services of special departments. In the general index he finds a listing of all laboratory studies and other procedures that the various diagnostic facilities programmed in the computer can perform. The situation is similar to that in a hospital or clinic where the physician must consult a laboratory manual listing the studies the laboratory will perform. If he considers it important to determine blood glucose, the computer might print out "90 mg per 100 ml." If he wishes a roentgenogram of the chest, the computer will print out the report as it would come to him directly from the radiology department. Although the results are given specifically, the index of questions and procedures is purposely so generalized that the physician must first determine in his own mind the questions he considers pertinent to the patient's problem and the laboratory procedure and diagnostic studies he wishes to order.

After the results of the requested diagnostic studies have been reported by the computer in response to specific requests, the final diagnosis is called for. The program might then challenge the examinee to support his diagnosis, drawing upon the information he has gathered about the patient, and thus simulating one of the features of many oral examinations. The computer-based examination may end at this point or it may continue in a sequential step-by-step manner to follow the patient's course for days, weeks, or months (Hubbard, 1971, pp. 101-103).

Scoring can be adjusted according to the sequence of the candidate's responses. And the patient's status can be continually estimated by the computer as a result of each new candidate decision.

Validity of Tests of Subject Matter and Tests of Ability to Apply Subject Matter

Five types of validity are commonly distinguished. Face validity refers to the logical relevance of item content to the subject matter tested, that is, whether items reflect course content. Content validity refers to the representativeness of content sampled by items, that is, whether all aspects of the course are represented. Concurrent validity compares scores with other measures of present (sometimes past) behavior, such as grades. Construct validity compares scores to a theoretical model or scores derived from it. Finally, predictive validity compares scores with measures of behavior subsequent to the test, such as subsequent job performance.

Nearly all examination procedures considered thus far have adequate face and content validity. Some of them also present evidence, not so convincing, of concurrent and construct validity. None present adequate evidence of predictive validity. Findings are summarized below for the following professions: teaching, law, engineering, medicine.

Teaching. Scores on the National Teacher Examination are used by some school districts in hiring and other personnel decisions. The examination has a more than thirty-year history and provides a weighted combination examination total score (WCET) in professional education and general education; in addition specialized subject matter examinations are offered in 24 areas. In their careful and current review of the National Teacher Examination, Quirk, Witten, and Weinberg (1973) summarize studies of concurrent validity. Correlations between the WCET and undergraduate GPA (N of 16 studies) ranged from .23 to .74, median of .55. Correlations between the WCET and student-teaching grades (N=2) were -.01 and -.04. Correlations between WCET and ratings by college teachers or principals during student-teaching (N=6) ranged from -.03 to .18, median of .05. Thus examination results show a modest relationship with grades but are unrelated to evaluations of clinical practice during training.

No studies of construct validity of the National Teacher Examination were located for that review. Studies of predictive validity rely on ratings by principals or supervisors during the first year (N=7) ranging from -.15 to .45, median .11 and after three years (N=1), .10. Thus the examination's validity has not been established.

There are two major explanations for these low correlations. The first is a rational scepticism that an examination designed to measure knowledge should also predict practice. The second is the well known unreliability of supervisor ratings due both to interjudge disagreement and to changes which ratees may be experiencing (Bray and Moses, 1972).

Until better research which shows clearer relationships with practice is produced, it is difficult to identify practical utility for the National Teacher Examination.

Law. Research has centered on the prediction of law school grades from the Law School Admission Test. Goolsby (1967) found a correlation of .60 between Law School Admission Test scores and law school grades with the Bar Examination scores. Although evidence is lacking, it is probable that the Bar Examination (particularly the Multistate Bar Examination) possesses high face and content validity, modest concurrent and construct validity, and little predictive validity. One critic of the Multistate Bar Examination contends that its approach is to "cover an entire subject with its penumbra of debatable cases and divergent views by exhorting the applicants to ignore refinements and pick the proper response by drawing upon that assemblage of 'majority' rules, 'traditional' rules, and 'trends' which he presumably carries in his head" (Pock, 1973, p. 67). To the extent that is true, the examination may correlate with the Law School Admission Test and with law school grades, but there is no reason to expect it to predict effective practice. I understand that a predictive validity study of the Multistate Bar Examination is in the planning stages, but thus far the problem of specifying criteria of effective law practice has not been solved.

Engineering. The National Commission of Engineering Examiners prepares a Fundamentals of Engineering Examination usually taken after graduation from an accredited school and the Principles of Practice Examination taken usually after four years of practice. The California Board of Registration for Professional Engineers has begun a predictive validity study of its examinations. Three hundred candidates randomly selected from those who took the Fundamentals Examination in April, 1971, will be followed through remaining steps of registration and into their careers. The only results thus far available are that the 300 adequately represent the 1971 examinees (Hoerger, 1972).

Medicine. Research assessing validity of the National Boards and of Specialty Boards is voluminous. The examinations are constructed to insure face and content validity. In a concurrent validity study, sophomore grades in 37 medical schools showed correlations with scores on Part I and Part II of the National Boards ranging from .63 to .84, mean .76. Correlations between senior grades and Part II ranged from .36 to .80, mean .68 (Hubbard, 1971).

Construct validity studies compare examination scores for groups of students at different points in their training. Part III scores were compared for students at the end of their third year in medical school and those near the end of their internship. Ninty percent of the third-year students had Part III scores below the mean of the intern scores. This finding was confirmed when the same third-year students were tested on an equivalent form of the examination at the end of their internship (Hubbard, 1971).

Residents in internal medicine took the 1969 written test of the American Board of Internal Medicine. Scores of third-year residents were higher than for first-year residents (Schumacher, 1973).

Levine, McGuire, and Nattress (1970) report similar findings for resident orthopedic surgeons. The latter study also provides evidence on concurrent validity. It evaluated the 1968 Certification Examination and the 1966 In-training Examination of the American Board of Orthopedic Surgeons and compared examination scores with supervisor ratings of residents. Scores in both examinations "are more closely related to supervisors' ratings of cognitive components of competence than to their ratings of skills and affect (Levine, McGuire, and Nattress, 1970, p. 76).

Concurrent validity of PMP's failed to receive support in a study by Goran et al. (1973). Clinic performance of 22 teams dealing with patients presenting symptoms of urinary tract infection was studied. A PMP was subsequently constructed from data on these 22 patients. Level of performance on the PMP was higher than in the clinic. PMP performance did not discriminate between poor, average, and excellent clinic performance. More extensive studies following this model are warranted.

The criterion problem has apparently dissuaded researchers from attempting predictive validity studies. They seem content with that state of affairs. "Because, as noted earlier, ultimate criteria of physician performance are not available, it is unlikely that any intermediate criterion will provide the 'critical experiment' to confirm or deny unequivocally the validity of National Board examinations. On the other hand, the validation studies that have been done, combined with the judgment of those who create the examinations, provide evidence that the examinations do measure what they are designed to measure" (Hubbard, 1971, p. 57).

I am less comfortable with that conclusion when I consider results of other investigations of criteria commonly used in concurrent validity research. Grades, for example, have never been shown to bear a notable relationship with post-school success. Wingard and Williamson (1973) located 27 studies (from 1955 to 1972) investigating professional school grades and career performance in medicine and other professions. Little or no relationship has been demonstrated (c.f. Hoyt, 1965; and Berg, 1970). Further, in that literature search, no training models were identified which had been shown to correlate positively with career performance. These findings suggest caution for the use of grades both in professional school admissions and as criteria for validity studies of certification examinations.

Since those who score low on these tests are thereby excluded from further training, validity studies suffer from the problem of "restriction of range." It is not possible to know how low scorers would have fared, because they are excluded. At least one study suggests that they may fare well. Forty-nine students were admitted to the University of Rochester School of Medicine and Dentistry despite Medical College Admission Test scores at least two standard deviations below the mean. Bartlett (1967) found that these students did not differ measurably from regularly admitted

students in medical school and in their later careers. A study is planned in Argentina where the Medical College Admission Test, translated into Spanish, will be taken by all incoming students. Since there the freshman year is the screening year, MCAT validity will be assessed against first year grades of all students (D'Costa, 1973).

It is likely that the tests discussed here are valuable devices for selecting persons for further academic study. But they are not thereby justified for determining candidates' readiness to practice. To the extent that effective performance in a profession involves other skills and competencies, these tests are not sufficient. The point is important because use of a test is not defensible in situations where it has not been validated. In a Supreme Court decision on a related issue, Justice Berger wrote in a unanimous opinion that only when a psychological test has been empirically related to job performance can it be used in employment decisions (Plotkin, 1972). The Bar Examination has already been challenged in the courts on the basis of alledged bias particularly with regard to race (American Civil Liberties Union, 1973).

It is likely that agencies using tests for certification decisions will be required to demonstrate the empirical relevance of those tests. The tests reviewed above would not fare well under such requirements except where they are used to select individuals for further academic study.

Some Suggestions for Examination Use

Tests used for certification purposes may be broadened in format to increase their usefulness. Gough has called for reducing the length of the Medical College Admission Test, in particular, and using the time saved for tests such as divergent thinking and cognitive flexibility (Gough, 1963). Rezler (undated) suggests greater use of problem solving tests even at the medical school admission level. Such variables are predominantly cognitive but would nevertheless broaden the range of cognitive abilities sampled.

The behavioral objectives model appears to have shaped current examinations. Attention to other models of educational evaluation might also broaden examinations. For example, Eisner describes two additional types of objectives. One is the expressive objective, "an outcome of an activity planned by the teacher or the student which is designed not to lead the student to a particular goal or form of behavior, but, rather, to forms of thinking-feeling-acting that are his own making" (Eisner, 1972, p. 580). His third type of objective, appropriately called Type III, is less open than the expressive but more open than the behavioral, where student terminal behavior is specified. Type III merely defines the problems specifically but does not identify solutions. The student task is to identify ingenious solutions. Heuristic concepts such as these may lead to more adequate examinations.

Another effective and appropriate use of present examinations is for self-assessment. Calls for self-assessment use have been made in law

(Bingaman, 1971), psychology (Tempone, 1971), and medicine (Hubbard, 1971; Knox, 1973). In this case only the professional receives test results for use in decisions about his own professional development. Thus he may interpret strengths and weaknesses relative to his own practice, may choose his own norm group for comparison, and may avoid the anxiety invariably attached to external evaluation. Although interest in self-assessment is growing, it has had little influence thus far on procedures for assessing readiness for practice.

VI. Assessing Simulated and Actual Job Performance

Few would be dissatisfied with making decisions on readiness for professional practice if the following data were available: a) a representative sample of job performance and b) a reliable system for rating that sample to insure comparability across candidates. Assessing actual job performance eliminates the need to infer from immediate criteria (grades) or intermediate criteria (supervisor ratings).

Yet most personnel selection work has not collected samples of behavior. Instead "signs" (scores on intelligence, aptitude, and personality tests) have been used, largely because of administrative convenience. The classic validity model requires that scores derived from these signs be correlated with the job performance criterion, usually supervisor ratings. The convention has been that predictor and criterion should somehow be different. Wernimont and Campbell (1968) provide the rationale for another model, the consistency model, in which predictors are as similar to the criteria as possible.

In a strong attack on "the testing movement," particularly intelligence testing, McClelland (1973) argues for a similar approach which he terms criterion sampling. Aptitude tests, for example, can be construed as providing criterion samples for school situations, situations in which their validity has indeed been demonstrated. But they provide criterion samples for virtually no life situations or job category. Thus new devices must be developed and new settings created for collecting work samples that are consistent with the criteria. Prediction of professional performance should rest on whether or not the candidate has exhibited such behavior in a comparable setting in the past. If he has never been in such a setting, one should be created or simulated for him. The practical difficulties of that procedure is undoubtedly a major reason that this approach has not been more used. But the consistency model has advantages as well. To the extent that what tests call for is consistent with what professional practice requires, faking is impossible; thus, test security is no longer an issue.

Performance-Based Teacher Education: An Example

A major effort to introduce consistency from training through certification to job performance is Performance-Based Teacher Education (PBTE). In PBTE the competence (readiness) of prospective teachers is judged by their direct demonstration of specified behaviors. "The teacher is asked to demonstrate his ability to perform certain critical teaching acts or to enable students to demonstrate certain specified abilities which they could not do before instruction." (Roth, 1973, p. 287). The American Association of Colleges for Teacher Education has strongly supported the development of PBTE and has stimulated extensive discussion of it in the literature.

Three of the essential characteristics of those teacher education programs which are performance-based are the following: "1) Competencies

(knowledge, skills, behaviors) to be demonstrated by the student [teacher trainee] are derived from explicit conceptions of teacher roles, stated so as to make possible assessment of a student's behavior in relation to specific competencies, and made public in advance. 2) Assessment of the student's competency uses his performance as the primary source of evidence, takes into account evidence of the student's knowledge relevant to planning for, analysing, interpreting, or evaluating situations or behavior, and strives for objectivity. 3) The student's rate of progress through the program is determined by demonstrated competency rather than by time or course completion" (Elam, 1971, pp. 6-7).

Some programs are beginning to operate under these guidelines and are struggling with the problems of measurement and criteria selection. Until competencies have been validated and reliable devices for their assessment have been developed, PBTE invites criticism. A United Federation of Teachers committee emphatically reports, "We will oppose any attempt to institute performance-based certification until validated research has been completed. Further, because there is no proven research on the subject of teacher competencies, we will oppose attempts to evaluate the performance of in-service teachers on this basis" (United Federation of Teachers, 1972).

A less self-interested critique of the concept of PBTE has been voiced by Broudy (1972). One of his major concerns is that mastery of theory for its own sake and as an informer of action is not compatible with the concept of PBTE. "...if performance of the specified task in a predetermined form is the criterion of success in teaching, then current programs in teacher preparation not only are unnecessarily abstract and theoretical, but perhaps otoise altogether. A program of apprenticeship training seems to be the only warranted investment of resources for the training of teachers. But once we arrive at this conclusion, it makes no sense to speak of 'professional' teachers as distinct from craftsmen, if professional means theory-guided practice with the practitioner possessing both the how and the why of the practice" (Broudy, 1972, pp. 11-12).

It will be some time before the potential of PBTE as a means for increasing consistency between test and criteria is known. (See the research bibliography, ERIC, 1972.) In the remainder of this chapter we shall examine several other attempts, relatively narrow ones, to assess performance in job-like settings. Behavior to be evaluated may be part of routine work or the work setting may be simulated. The simulation may or may not involve other people. Illustrations of these three categories are given below.

Non-interpersonal Simulations

A non-interpersonal simulation is one in which the candidate deals with materials and equipment rather than people. Data are presented in a more realistic context than merely as part of a test booklet or on the display panel of a computer.

One example is an attempt to assess empathy reported by Campbell, Kagen, and Krathwohl (1971). The candidate views video-taped exerpts from actual

counseling sessions. He is then presented with a multiple-choice question presenting three responses which the client might make next. That is, the candidate is to choose a response that demonstrates empathy for the client. The 89-item test was shown to be adequately reliable. Modest concurrent and construct validity has been established; there is no evidence on predictive validity. Such a procedure deserves further work because, although it samples only what the candidate says he would do (or say), the taped image increases realism.

More realistic, yet still not involving other persons, is the in-basket test. Developed for assessing managerial talent, in-basket simulates the variety of tasks a manager faces in the particular position being recruited. The candidate is asked to deal during a specified time period with a number of items (memos, letters, bills, etc.) as a good manager would. Each decision may be scored for content and for style. In a cross-validation study, Meyer (1970) found that scores bear a modest relationship to supervisor ratings and result in better predictions than a paper-and-pencil test. Little predictive validity research has been done on the test however, probably because of its high face validity. Validity studies should be well worth the effort.

Interpersonal Simulations

The in-basket technique in combination with other individual and group tasks has been employed by "assessment centers" in their management screening program. An assessment center uses techniques reminiscent of the large personality research projects described above. It attempts to simulate situations the candidate would face if he were promoted. "Groups of men pass through a series of standardized exercises such as management games, in-basket tests, and leaderless discussion sessions, while the assessors observe their behavior closely. The assessors discuss each candidate's performance separately and then generate a comprehensive report on each candidate which management can combine with current performance information as it sees fit. As well as identifying the men most likely to succeed, the assessment reports spell out the individual deficiencies of each candidate and suggest guidelines for management to use in developing him" (Byham, 1970, p. 151). In 1970 A.T.&T. alone operated 50 centers with 10,000 candidates per year.

These centers are expensive, probably more than $500 per candidate, but do yield more accurate prediction than traditional methods. Byham cites six reasons for their success: "The exercises used are designed to bring out the specific skills and aptitudes needed in the position(s) for which a group of candidates is being assessed. Since the exercises are standardized, assessors evaluate the candidates under relatively constant conditions and thus are able to make valid comparative judgments. The assessors usually do not know the candidates personally: so, being emotionally disengaged, they are unbiased. The assessors are shielded from the many interruptions of normal working conditions and can pay full attention to the candidates' behavior in the exercises. The procedures focus their attention

on the primary kinds of behavior they ought to observe in evaluating a promotion candidate. They have been trained to observe and evaluate these kinds of behaviors" (1970, p. 151). Many candidates show great involvement in these tasks, as they would on the job. Wollowick and McNamara (1969) in an assessment center study find support for combining data statistically rather than subjectively. The criterion was predicted at .62 by statistical means and only at .37 by subjective (clinical) means.

Nevertheless, assessment centers themselves deserve more careful assessment particularly in view of their high cost. Data showing that prediction accuracy is increased when situational measures are combined with paper-and-pencil tests uses statistical significance as criterion. If practical significance is considered, according to Wilson and Tatge (1973), superiority of the assessment center is unsubstantiated.

Persistent use of the bedside interview for evaluation of physicians has not received research support (Hubbard, 1971), as noted above. Simulated patient interviews are still being researched, however, and may prove more effective if improved devices for rating the interviews are developed (Lamont and Hennen, 1972; Levine and McGuire, 1970).

Simulations for the study of physician problem solving strategies have been developed by Norman Kagen and a group at Michigan State University (1971). Their thoughtful use of video-tape technology is especially noteworthy. The simulation room is decorated like a physician's office. The "patient" is a drama school actor who has been given detailed information about the case. The interaction is recorded by two ceiling mounted cameras. After the physician has completed the patient interview, the actor is replaced by an assistant who serves as a "data bank" and dispenses whatever physical findings and laboratory test results the physician requests. (In another case, an actress simulates both history and symptoms, including emotional problems.)

After the work-up is completed, the physician views the tape and a research assistant prompts him to recall what he was thinking and feeling during the original interaction. This interview is also taped for subsequent analysis. In other uses of the "interpersonal process recall" procedure, subjects have been wired for physiological measures (respiration, galvanic skin response, etc.). A split-screen displays both the subject's interpersonal responses during the simulated interaction and the graphical record of his physiological responses. The subject's subsequent review of that tape with a research assistant is also recorded thus providing a final tape showing the subject being interviewed while viewing a tape of an interaction and of his physiological responses during that interaction (Kagen, 1973). The considerable amount of resulting data may then be analysed to compare performance on such dimensions as the number and type of questions necessary for the physician to formulate the correct diagnosis. Such simulations provide standardization and control of the information given to the examinee while also introducing the complexity of human interaction.

On the Job Work Samples

Evaluations of on the job performance have typically been little more than supervisor global ratings of performance. More useful are evaluations of carefully delimited situations using reliable data recording and scoring procedures. For example, Kopta (1971) investigated the operative skills of 40 residents in orthopedic surgery. He derived a rating sheet which provided reliable data and demonstrated construct validity. He contends,"...if the operating room is to continue to serve as a major setting for teaching and learning, it should also serve as a major setting for evaluation" (Kopta, 1971, p. 302).

An aid in performance assessment may be the "clinical algorithm" developed for the training of physicians' assistants (Sox, Sox, and Tompkins, 1973). Algorithms were developed for medical problems (e.g., cough, fever, shortness of breath) in order to provide unambiguous step-by-step instructions for dealing with the complaint. The algorithm is used in instruction, a medical record form corresponding to the algorithm is used by the assistant during a work-up, and a computer assesses the assistant's accuracy in completing the form according to the algorithm. Where carefully circumscribed critical performance situations can be identified, the algorithm should be a useful device to assist observation and evaluation as well as instruction.

A final example of on the job observation of criterion behavior is the effective use of "minicourses" for teacher training.

> In taking the minicourse, the trainee first views an instructional film in which one to three specific teaching skills are described and illustrated with examples from various classrooms. This instructional film is followed by a model film in which the trainee sees a model teacher fitting these skills into a regular classroom lesson. As part of viewing the model film, the trainee is usually called upon to identify each skill as the model teacher uses it. After viewing the instructional and model films, the trainee receives further information on the specific skills by studying a teacher handbook. He then prepares a microteach lesson designed to give him practice in using the skills. Microteach lessons, as used in Minicourse 1, are typically 10-20 minutes in length and are taught to a group of from 4 to 10 pupils taken from the inservice teacher's regular classroom. The lesson is recorded on video-tape. Immediately after the lesson, the trainee replays the recording and evaluates his use of the teaching skills employing evaluation forms provided in the teacher handbook. Then, based on his evaluation, he replans the lesson and reteaches it to another small group of pupils from his regular classroom, again recording the lesson on video-tape and evaluating it using additional evaluation forms provided in the handbook. Thus, the minicourse instructional model contains three main elements: (a) the

instructional and model films, (b) the microteach and reteach lessons, and (c) the videotape replay and self-evaluation (Borg, 1972, p. 573).

Field tests of the minicourse have collected data in the teacher's regular classroom although lessons are typically practiced with small groups. Forty teachers in twelve schools presented 20-minute lessons for video-taping before and after the course. Teacher talk occupied about 52% of class time before training; training reduced that by nearly half, to 28%. Side benefits included more pupil interaction and less teacher restriction of discussion. A second skill, although obviously related to the first, is the use of questions that call for longer pupil responses. Training resulted in the near doubling of the average word count of pupil responses and significant decrease in the incidence of one-word pupil replies. As a final skill example, teacher questions were categorized as to whether they called for mere recall of information or for higher cognitive processes. Before training, only 37% of the questions called for higher cognitive processes; after training 52% did so, presumably raising the quality of pupil talk. Five months after completing the course, another set of tapes made for follow-up analysis showed that all but one skill remained at its post-course level, three skills were actually strengthened in the interim. A second follow-up three years after the training found teacher performance still superior to pre-training performance on eight of the ten skills scored. These results are strong evidence not only for the effectiveness of the minicourse procedures, but also for the specificity and reliability of the observations made of on the job criterion performance.

For obvious reasons attempts to assess performance directly under either simulated or on the job conditions have been most frequent in areas such as languages, art, music, vehicle operation, secretarial skills, industrial arts, and agriculture. If performance tests are to be used in assessing readiness for professional practice, care must be taken with regard both to their comprehensiveness, i.e. the range of aspects of the profession sampled, and their fidelity, i.e. the degree of accuracy with which those aspects are sampled (Fitzpatrick and Morrison, 1971). Use of adequately developed performance tests which complement paper-and-pencil tests of cognitive abilities, are essential if assessment is to have satisfactory predictive validity.

VII. Summary and Conclusions

Readiness for professional practice is defined and assessed by those agencies with licensing and certifying power (educational institutions, professional associations, and government agencies).

Criteria for effective practice should be derived by analysing the responsibilities and duties of the professionals in question. Identifying such responsibilities and duties may involve data from groups of colleagues and clients and from institutional records. Diaries, trained observers, and critical incident surveys are often used. Other criteria may be identified through reviews of research literature, particularly in teaching, which have identified professional behaviors and characteristics found to be correlated with client outcomes. Unfortunately, criteria identified through surveys or research reviews have limited validity in that they reflect a particular Zeitgeist. Use of a culturally universal developmental framework such as Kohlberg's moral development scheme may reduce the subjectivity in criterion selection.

Once criteria are selected, they must be operationalized so that reliable comparisons across candidates may be made. The operations may take the form of work samples, in which case the predictor and the criterion are consistent. Or the predictor may be dissimilar in form from the criterion. Frequently used predictors include tests of personality characteristics, tests of subject matter knowledge, and tests of subject matter application. When tests are used, evidence is required that scores on the predictor do in fact correlate with professional effectiveness. Consequently, professional effectiveness must then be independently operationalized as a criterion measure.

Reviews of several major personnel selection programs using personality tests suggest questionable utility, given their high cost. Effectiveness is increased when findings are statistically combined for decision making rather than when clinical judgments are used.

Tests of subject matter knowledge are the most widely used method for determining readiness. In most professions they are multiple-choice in form, although essay bar examinations are used in most states. Little evidence for the usefulness of oral examinations is available, and their use is declining. Tests of ability to apply subject matter are usually administered as part of these examinations and present the candidate with a novel situation which he has not previously encountered in his training. In medicine, for example, items may take the form of patient management problems.

The most widely used examinations, such as those in teaching, law, engineering, and medicine, are constructed according to principles which assure appropriate item difficulty and discrimination as well as high reliability and face validity. Some evidence on concurrent and construct validity is available, but the predictive validity of the tests has not

been established. That is, there has not been empirical demonstration that scores on the test are highly correlated with subsequent professional practice. Indeed, since the tasks presented by the tests and those encountered in practice differ greatly, it seems unlikely that validity will be demonstrated even if the formidable criterion problems are solved. Since both scientific and professional ethics require that a selection device have demonstrated validity for its use, continued emphasis on such tests is likely to be controversial.

If decisions on readiness can be based on a representative sample of job performance, reliably rated, the question of validity is answered: the predictor and the criterion are nearly identical. Performance-Based Teacher Education is an attempt to introduce consistency from training through certification to job performance. But PBTE must nevertheless demonstrate that the performances it requires are valid (that is, are related to some "good thing" such as student learning) and are being accurately measured. Further, it must satisfy critics who question the implication that an array of observable behaviors adequately comprises a definition of professional effectiveness.

Among procedures for gathering work samples are simulated physician-patient interviews, leaderless group exercises, empirically derived skill rating systems, and the use of trained observers. These techniques are most useful when only carefully delimited situations are assessed and when data from different situations are combined through statistical rather than clinical procedures.

Several conclusions seem apparent. First, criterion analysis should emphasize many discrete behaviors and characteristics rather than global definitions. Second, measures of these characteristics (predictors) should be as similar to the criterion itself as possible. Third, multiple assessment devices should be used so that no type is overemphasized. Measures of subject matter knowledge, in particular, should be less heavily weighted. Fourth, data should be combined by statistical rather than clinical methods. Fifth, data should not be used for decision making until longitudinal studies demonstrate adequate predictive validity.

Finally, let us imagine a hypothetical assessment program with high predictive validity. One consequence of such a program would be an intensification of the adversary relationship between the candidate and the certifying professionals. The adversary relationship derives from the fact that the latter has something, i.e. certification, that the former wants. Presently, because assessment is imperfect, many candidates, including some of the most creative, may be expected to try to subvert it. If accuracy were increased, even more subversion attempts would occur. That is an especially unfortunate state of affairs since the most accurate decisions are likely to result from open collaboration between candidate and examiner. Collaboration would sharpen the candidate's skills of self-evaluation, and a professional skilled in self-evaluation is far more valuable than one who has merely outwitted a professional evaluator.

Thus increased assessment validity has at least two negative consequences which may be self-defeating. Examinee defensiveness is likely to increase, a) eliminating any chance of open, collaborative communication with the examinee and b) increasing cheating. Second, an increase in examinee anxiety may interfere with performance thereby increasing the number of false negatives.

The dilemma of assessing readiness for professional practice in a competitive, adversary context is therefore the following: the irresponsibility of procedures with undemonstrated predictive validity on the one hand and the self-defeating nature of procedures with high predictive validity on the other. The imperative for assessment is to focus on predictive validity while moving toward more open, collaborative evaluation.

References

Adams, F. H. The review and revision of certification procedures in pediatric cardiology. Journal of Medical Education, 1972, 47, 796-805.

Adams, H. B. Performance evaluation in ministry. Theological Education, 1971, 7, 102-108.

Adams, T. H., et al. Law schools and bar admission requirements: A review of legal education in the United States--Fall 1972. Chicago: American Bar Association Section of Legal Education and Admissions to the Bar, 1972.

American Civil Liberties Union. Bar exam bias. Civil Liberties, 1973, July (No. 297), 8.

American Dental Association. The dental admission testing program. Journal of Dental Education, 1971, 82, 1051-1056.

Asher, J. Opposition wins latest round in fight over New York licensing bill. American Psychological Association Monitor, 1973, 4, 1;12.

Bartlett, J. W. Medical school and career performances of medical students with low Medical College Admission Test scores. Journal of Medical Education, 1967, 42, 231-237.

Becker, H. S. The nature of a profession. In N. B. Henry (Ed.), Education for the professions. The sixty-first yearbook of the National Society for the Study of Education. Part II. Chicago: University of Chicago Press, 1962. Pp. 27-46.

Berg, I. Education and jobs: The great training robbery. New York: Praeger, 1970.

Bingaman, C. C. Illinois Bar Examination: Time for a 75-year overhaul. Illinois Bar Journal, 1971, 60, 129-139.

Blum, M. J., and Fitzpatrick, R. Critical performance requirements for orthopedic surgery. Chicago: University of Illinois College of Medicine, February, 1965.

Borg, W. R. The minicourse as a vehicle for changing teacher behavior: A three-year follow-up. Journal of Educational Psychology, 1972, 63, 572-579.

Bray, W., and Moses, J. L. Personnel selection. Annual Review of Psychology, 1972, 23, 545-576.

Briggs-Myers, I., and Davis, J. A. Relation of medical students' psychological type to their specialities twelve years later. Princeton, N. J.: Educational Testing Service Research Memorandum, RM-64-15, 1964.

Briner, L. A. The new Presbyterian system of evaluating candidates for ordination. Theological Education, 1971, 7, 92-101.

Brody, B. L., and Stokes, J. Use of professional time by internists and general practitioners in group and solo practice. Annals of Internal Medicine, 1970, 73, 741-749.

Broudy, H. S. A critique of performance-based teacher education. Washington, D. C.: American Association of Colleges for Teacher Education, 1972. (ERIC number ED 063 274)

Brown, C. R., Jr., and Uhl, H.S.M Mandatory continuing education--Sense or nonsense? Journal of the American Medical Association, 1970, 213, 1660-1668.

Byham, W. C. Assessment centers for spotting future managers. Harvard Business Review, 1970, 48, 150-164.

Campbell, R. J., Kagan, N., and Krathwohl, D. R. The development and validation of a scale to measure affective sensitivity (empathy). Journal of Counseling Psychology, 1971, 18, 407-412.

Covington, J. E. Multistate Bar Examination--A new approach. Arkansas Law Review, 1972, 26, 153-168.

D'Costa, A. Concerning MCAT validity and related questions; Report on research based on current admissions testing problems. Medical College Admissions Assessment Program Report, 1973, 1, 5-8.

Dumont, M. P. The changing face of professionalism. Social Policy, 1970, 1, 26-31.

Eckler, J., and Covington, J. E. The new Multistate Bar Examination. American Bar Association Journal, 1971, 57, 1117-1120.

Eisner, E. W. Emerging models for educational evaluation. School Review, 1972, 80, 573-590.

Elam, S. Performance-based teacher education. What is the state of the art? Washington, D. C.. American Association of Colleges for Teacher Education Committee on Performance-Based Teacher Education, 1971. (ERIC number ED 058 166)

ERIC Clearinghouse on Teacher Education. Performance-based teacher education: An annotated bibliography. Washington, D. C.: American Association of Colleges for Teacher Education, 1972. (ERIC number ED 065 477)

Ezekiel, R. S. The personal future and Peace Corps competence. Journal of Personality and Social Psychology, 1968, 8 (2, Pt. 2), 1-26.

Fiske, D. W., and Pearson, P. H. Theory and techniques of personality measurement. Annual Review of Psychology, 1970, 21, 49-86.

Fitzpatrick, R., and Morrison, E. J. Performance and product evaluation. In R. L. Thorndike (Ed.), Educational measurement.(2nd ed.) Washington, D. C.: American Council on Education, 1971. Pp. 237-270.

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Flavell, J. H. The developmental psychology of Jean Piaget. New York: Van Nostrand, 1963.

Foster, J. T., Abrahamson, S., Lass, S., Girard, R., and Garris, R. Analysis of an oral examination used in specialty board certification. Journal of Medical Education, 1969, 44, 951-954.

Goldberg, L. R. Reliability of Peace Corps selection boards: A study of interjudge agreement before and after board discussions. Journal of Applied Psychology, 1966, 50, 400-408.

Goolsby, T. M. A study of the criteria for legal education and admission to the bar. Journal of Legal Education, 1967, 20, 175-177.

Goran, M. J., Williamson, J. W., and Gonnella, J. S. The validity of patient management problems. Journal of Medical Education, 1973, 48, 171-177.

Gough, H. G., Hall, W. B., and Harris, R. E. Admissions procedures as forecasters of performance in medical training. Journal of Medical Education, 1963, 38, 983-998.

Greising, R. (Ed.) A candidate's guide to the Blue Ribbon Examination to Practice the Profession of Pharmacy. Chicago: National Association of Boards of Pharmacy, 1972.

Gross, R., and Osterman, P. (Eds.) The new professionals. New York: Simon and Schuster, 1972.

Guion, R. M. Personnel selection. Annual Review of Psychology, 1967, 18, 191-216.

Hale, R. Report of summer research. Dayton, Ohio: American Association of Theological Schools, undated.

Harris, J. G., Jr. A science of the South Pacific: Analysis of the character structure of the Peace Corps volunteer. American Psychologist, 1973, 28, 232-247.

Harvey, O. J., Hunt, D. E., and Schroder, H. M. Conceptual systems and personality organization. New York: Wiley, 1961.

Hoerger, H. J. Career 300--First interim report. Sacramento, California: Board of Registration for Professional Engineers, 1972.

Hoffman, J. J., and Nattress, L. W., Jr. Assessment of problem-solving ability: A proposal. Chicago: Natresources, Inc., undated.

Hubbard, J. P. Measuring medical education: The tests and test procedures of the National Board of Medical Examiners. Philadelphia, Pennsylvania: Lea and Febiger, 1971.

Hunt, D. E. Adaptability in interpersonal communication among training agents. The Merrill-Palmer Quarterly of Behavior and Development, 1970, 16, 325-344.

Hunt, D. E. Matching models in education: The coordination of teaching methods with student characteristics, Monograph series no. 10. Ontario, Canada: Ontario Institute for Studies in Education, 1971.

Hoyt, D. P. The relationship between college grades and adult achievement: A review of the literature. American College Testing Program Research Report, 1965, No. 7.

Ivey, A. E., and Rollin, S. A. A behavioral objectives curriculum in human relations: A commitment to intentionality. Journal of Teacher Education, 1972, 23, 161-165.

Jackson, P. W. The difference teachers make. In United States Department of Health, Education and Welfare, How Teachers Make a Difference. Washington, D. C.: United States Government Printing Office, 1971. (OE-58044) Pp. 21-31.

Jones, R. R. The Peace Corps overseas: Some first steps toward description and selection, Technical Report 8, No. 3. Oregon Research Institute, 1968. (a)

Jones, R. R. The validity of the Full Field Background Report in Peace Corps selection, Research Monograph 8, No. 1. Oregon Research Institute, 1968. (b)

Jones, R. R. Selection and overseas experiences of Peace Corps Volunteers, Final Report. Oregon Research Institute, August, 1969.

Kagan, N. Can technology help us toward reliability in influencing human interactions. Educational Technology, 1973, 13 (2), 44-51.

Kagan, N., et al. Methods for the study of medical inquiry. East Lansing, Michigan: Office of Medical Education Research and Development, Michigan State University, 1971.

Kelly, E. L., and Fiske, D. W. The prediction of performance in clinical psychology. Ann Arbor: University of Michigan Press, 1951.

Kelly, E. L., and Goldberg, L. R. Correlates of later performance and specialization in psychology: A followup of the trainees assessed in the VA selection research project. Psychological Monographs, 1959, 73 (12), 1-32.

Knox, A. B. Life long self directed education. Urbana, Illinois: University of Illinois College of Education, 1973.

Koewing, J. R. Seekers and colleagues: An interdisciplinary approach for continuing professional education. Chicago, Illinois: Natresources, Inc., 1972.

Kohlberg, L. The contribution of developmental psychology to education--Examples from moral education. Educational Psychologist, 1973, 10, 2-14.

Kohlberg, L., and Turiel, E. Moral development and moral education. In G. S. Lesser (Ed.), Psychology and educational practice. Glenview, Illinois: Scott, Foresman, 1971. Pp. 410-465.

Kopta, J. A. An approach to the evaluation of operative skills. Surgery, 1971, 70, 297-302.

Lamont, C. T., and Hennen, B. K. E. The use of simulated patients in a certification examination in family medicine. Journal of Medical Education, 1972, 47, 789-795.

Levine, H. G., and McGuire, C. H. The use of role-playing to evaluate affective skills in medicine. Journal of Medical Education, 1970, 45, 700-705.

Levine, H. G., McGuire, C. H., and Nattress, L. W., Jr. The validity of multiple choice achievement tests as measures of competence in medicine. American Educational Research Journal, 1970, 7, 69-82.

Mackert, M. C. Bar examinations: good moral character, and political inquiry. Wisconsin Law Review, 1970, 1970, 471-494.

MacKinnon, D. W., and Crutchfield, R. S., Barron, F., Block, J., Gough, H. G., and Harris, R. E. An assessment study of Air Force officers: Part I. Design of the study and description of the variables. Lackland Air Base, Texas: Wright Air Development Center, Personnel Laboratory, Technical Report No. 91 (I), April, 1958.

McClelland, D. C. Testing for competence rather than for "intelligence." American Psychologist, 1973, 28, 1-14.

McGuire, C. H. The oral examination as a measure of professional competence. Journal of Medical Education, 1966, 41, 267-274.

Meyer, H. H. The validity of the in-basket test as a measure of managerial performance. Personnel Psychology, 1970, 23, 297-307.

Miller, G. E., McGuire, C. H., and Larson, C. B. The orthopaedic training study. Bulletin of the American Academy of Orthopaedic Surgeons, 1965, 13, 8-11.

National Council for Accreditation of Teacher Education. Standards for accreditation of teacher education: the accreditation of basic and advanced preparation programs for professional school personnel. Washington, D. C.: Author, 1970.

National Board of Medical Examiners. National Board highlights--1972. The National Board Examiner, 1973, 20 (6), all.

National League for Nursing. A validation study of the National League for Nursing Pre-Nursing and Guidance Examination (and related studies emerging from data gathered for the validation study). New York: Author, 1970. Publication no. 17-1390.

Nattress, L. W., Jr. Oral tests--their reliability, validity, and objectivity: A survey of the literature. Chicago: Natresources, Inc., 1964.

Nattress, L. W., Jr. The branching-type written simulative exercises. Chicago: Natresources, Inc., 1970.

Office of Strategic Services Assessment Staff. Assessment of men. New York: Rinehart, 1948.

Packer, H. L., and Ehrlich, T. New directions in legal education. New York: McGraw-Hill, 1972.

Plotkin, L. Coal handling, steamfitting, psychology, and law. American Psychologist, 1972, 27, 202-204.

Pock, M. A. The case against the objective Multistate Bar Examination. Journal of Legal Education, 1973, 25, 66-71.

Price, P. B., Lewis, E. G., Loughmiller, G. C., Nelson, D. E., Murray, S. L., and Taylor, C. W. Attributes of a good practicing physician. Journal of Medical Education, 1971, 46, 229-237.

Quirk, T. J., Witten, B. J., and Weinberg, S. F. Review of studies of the concurrent and predictive validity of the National Teacher Examinations. Review of Educational Research, 1973, 43, 89-113.

Rezler, A. G. A position paper presenting recommendations for development of an admissions assessment program for medical colleges. Chicago: University of Illinois Medical College, undated.

Rogers, C. R. Some new challenges. American Psychologist, 1973, 28, 379-387.

Rosenshine, B., and Furst, N. Research on teacher performance criteria. In B. O. Smith (Ed.), Research in teacher education: a symposium. Englewood Cliffs, N. J.: Prentice-Hall, 1971. Pp. 37-72.

Roth, R. A. Certifying teachers: an overhaul is underway. The Clearing House, 1973, 47, 287-291.

Schein, E. H. Professional education: Some new directions. New York: McGraw-Hill, 1972.

Schumacher, C. F. Validation of the American Board of Internal Medicine Written Examination: A study of the examination as a measure of achievement in graduate medical education. Annals of Internal Medicine, 1973, 78, 131-135.

Sedlacek, W. E., and Nattress, L. W., Jr. A technique for determining the validity of patient management problems. Journal of Medical Education, 1972, 47, 263-266.

Shafer, W. G. How a National Board Examination is constructed. Journal of Dental Education, 1968, 32, 188-190.

Smith, M. B. Explorations in competence: A study of Peace Corps teachers in Ghana. American Psychologist, 1966, 21, 555-566.

Sox, H. C., Sox, C. H., and Tompkins, R. K. The training of physician's assistants. New England Journal of Medicine, 1973, 288, 818-824.

Super, D. E., and Crites, J. O. Appraising vocational fitness. (2nd ed.) New York: Harper and Brothers, 1962.

Tempone, V. J. The American Board of Professional Psychology's examination procedure: A time for change. Professional Psychology, 1971, 2, 177-182.

Thomas, M. C. Report of summer research. Dayton, Ohio: American Association of Theological Schools, undated.

Thorndike, R. L. Personnel selection: Test and measurement techniques. New York: Wiley, 1949.

Turiel, E. Stage transition in moral development. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand-McNally, 1973. Pp. 732-758.

United Federation of Teachers. UFT committee reports. PBTE Newsletter (Albany), 1972, 1 (1), 5.

United States Department of Health, Education and Welfare. Report on licensure and related health personnel credentialing. Washington, D. C.: DHEW Publication No. (HSM) 72-11, June, 1971.

Wernimont, P. H., and Campbell, J. P. Signs, samples, and criteria. Journal of Applied Psychology, 1968, 52, 372-376.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Massachusetts: Addison-Wesley, 1973.

Williamson, J. W. Assessing clinical judgment. Journal of Medical Education, 1965, 40, 180-187.

Wilson, J. E., and Tatge, W. A. Assessment centers--Further assessment needed? Personnel Journal, 1973, 52, 172-179.

Wingard, J. R., and Williamson, J. W. Grades as predictors of physician's career performance: An evaluative literature review. Journal of Medical Education, 1973, 48, 311-322.

Winter, J. A., Mills, E. W., and Hendrick, P. S., et al. Clergy in action training: A research report. New York: IDOC-North America, Inc., 1971.

Wollowick, H. B., and McNamara, W. J. Relationship of the components of an assessment center to management success. Journal of Applied Psychology, 1969, 53, 348-352.

APPENDICES

Procedure

Materials were gathered for this report during April, May, and June of 1973. Three procedures were followed. Retrieval was initiated from the following computerized data banks. ERIC, Dissertation Abstracts, and Science Information Exchange. Selected journals were searched for the last five to ten years (see attached list). Finally, contacts were made with more than 25 agencies and professional associations, usually including extended telephone conversations (see attached list).

Consequently, a great many persons have contributed the range of materials available for study. Nevertheless, this survey does not pretend to be exhaustive. The modest aim is to adequately illustrate the actual and potentially useful activities related to assessing readiness for professional practice.

Resource Journals

Journal	Years Searched
American Bar Association Journal	1967-71, 1973.
American Journal of Nursing	1963-1972
American Journal of Occupational Therapy	1963-1972
American Journal of Pharmaceutical Education	1963-1972
American Psychologist	1967-1973
Journal for the Scientific Study of Religion	1965-1970
Journal of Dental Education	1963-1972
Journal of Educational Psychology	1963-1973
Journal of Educational Research	1972-1973
Journal of Legal Education	1948-1970
Journal of Medical Education	1969-1973
Journal of Teacher Education	1967-1973
Journal of the American Medical Association	1963-1973
Professional Psychology	1969-1973
Religious Education	1967-1972
Theological Education	1967-1972

Major Associations and Agencies
Contacted for this Report

Architecture:

American Institute of Architects
1735 New York Avenue
Washington, D. C. 20006

National Architectural Accrediting Board
1785 Massachusetts Avenue, N. W.
Washington, D. C. 20036

Dentistry:

American Dental Association
211 East Chicago Avenue
Chicago, Illinois 60611

Dietetics:

American Dietetic Association
620 North Michigan Avenue
Chicago, Illinois 60611

Education:

American Association of Colleges for Teacher Education
One DuPont Circle
Washington, D. C. 20036

National Council for Accreditation of Teacher Education
1750 Pennsylvania Avenue, N. W.
Washington, D. C. 20006

Multi-State Consortium on Performance-Based Teacher Education
New York State Education Department
Albany, New York

Engineering:

Engineers' Council for Professional Development
345 East 47th Street
New York, New York 10017

Associations and Agencies (Cont'd)

National Council of Engineering Examiners
Box 752
Clemson, South Carolina 29631

Board of Registration for Professional Engineers
1021 "O" Street, Room A-102
Sacramento, California 95814

Law:

American Bar Foundation
1155 East 60th Street
Chicago, Illinois 60637

Association of American Law Schools
One DuPont Circle, N. W., Suite 370
Washington, D. C. 20036

American Bar Association
Section of Legal Education and Admissions to the Bar
1155 East 60th Street
Chicago, Illinois 60637

Medicine:

Center for Educational Development
University of Illinois Medical Center
1737 West Park
Chicago, Illinois

National Board of Medical Examiners
3930 Chestnut
Philadelphia, Pennsylvania 19104

American Medical Association
Council on Medical Education
535 North Dearborn Street
Chicago, Illinois 60610

Association of American Medical Colleges
One DuPont Circle, N. W., Suite 200
Washington, D.C. 20036

Associations and Agencies (Cont'd)

School of Medical Sciences
Pacific Medical Center
P. O. Box 7999
San Francisco, California 94120

Ministry:

American Association of Theological Schools
534 Third National Building
Dayton, Ohio 45402

Nursing:

National Association for Practical Nurse Education and Service, Inc.
1465 Broadway
New York, New York 10036

National League for Nursing
Department of Baccalaureate and Higher Degree Programs
10 Columbus Circle
New York, New York 10019

Optometry:

American Optometric Association
Council of Optometric Education
7000 Chippewa Street
St. Louis, Missouri 63119

Pharmacy:

American Council on Pharmaceutical Education
77 West Washington Street
Chicago, Illinois 60602

National Association of Boards of Pharmacy
77 West Washington Street
Chicago, Illinois 60602

Associations and Agencies (Cont'd)

Psychology:

American Psychological Association
Education and Training Board
1200 Seventeenth Street, N. W.
Washington, D. C. 20036

American Association of State Psychology Boards
School Administration Building
253 Prairie Avenue
Cheyenne, Wyoming 82001

Social Work:

Council on Social Work Education
Division of Education Standards and Accreditation
345 East 46th Street
New York, New York 10017

National Association of Social Workers
Southern Building, 6th Floor
15th and H Street, N. W.
Washington, D. C. 20005

Miscellaneous:

Natresources, Inc.
520 North Michigan Avenue
Chicago, Illinois 60611